Annals of Clinical Epidemiology 2024;6(4):77–86

Annals of Clinical Epidemiology

ORIGINAL ARTICLE

Bayesian Latent Class Models for Evaluating the Validity of Claim-based Definitions of Disease Outcomes

Satoshi Uno[1,2,3], Toshiro Tango[2]

ABSTRACT

BACKGROUND

Large electronic databases have been widely used in recent years; however, they can be susceptible to bias due to incomplete information. To address this, validation studies have been conducted to assess the accuracy of disease diagnoses defined in databases. However, such studies may be constrained by potential misclassification in references and the interdependence between diagnoses from the same data source.

METHODS

This study employs latent class modeling with Bayesian inference to estimate the sensitivity, specificity, and positive/negative predictive values of different diagnostic definitions. Four models are defined with/without assumptions of the gold standard and conditional independence, and then compared with breast cancer study data as a motivating example. Additionally, simulations that generated data under various true values are used to compare the performance of each model with bias, Pearson-type goodness-of-fit statistics, and widely applicable information criterion.

RESULTS

The model assuming conditional dependence and non-gold standard references exhibited the best predictive performance among the four models in the motivating example data analysis. The disease prevalence was slightly higher than that in previous findings, and the sensitivities were significantly lower than those of the other models. Additionally, bias evaluation showed that the Bayesian models with more assumptions and the frequentist model performed better under the true value conditions. The Bayesian model with fewer assumptions performed well in terms of goodness of fit and widely applicable information criteria.

CONCLUSIONS

The current assessments of outcome validation can introduce bias. The proposed approach can be adopted broadly as a valuable method for validation studies.

KEY WORDS

latent class model, Bayesian, outcome validation, epidemiology

[1] The Graduate University for Advanced Studies (SOKENDAI), Tachikawa-Shi, Tokyo, Japan
[2] Center of Medical Statistics, Minato-Ku, Tokyo, Japan
[3] Astellas Pharma Inc., Chuo-Ku, Tokyo, Japan

Corresponding author: Satoshi Uno
The Graduate University for Advanced Studies (SOKENDAI), 10-3, Midori-Cho, Tachikawa-Shi, Tokyo 190-8562, Japan
E-mail: uno.satoshi@ism.ac.jp

Received: December 26, 2023
Accepted: May 31, 2024
J-STAGE Advance published date: July 18, 2024
Published: October 1, 2024

DOI: https://doi.org/10.37737/ace.24012

INTRODUCTION

In recent years, large electronic databases (DB) have become prominent in epidemiological studies, with DBs of administrative claims and electronic medical records commonly used[1]. Although the evidence derived from these DB studies has already been widely used for public health, clinical practice, and policymaking, the potential for information bias poses a significant concern[2]. Notably, misclassification of disease occurrence, drug usage, and disease severity are well-recognized issues, as many DBs lack detailed clinical and pathological information[3]. To address these concerns, validation studies assessing the accuracy of coded algorithms to define the aforementioned factors against a reference have become critical components in demonstrating the validity of using DB studies for epidemiologic research[3]. A typical validation study is conducted in the following manner. First, some possible diagnostic definitions are developed to identify outcome events from electronic databases (DB diagnoses), such as diagnosis, surgery, laboratory tests, treatment, and their combinations. Second, the gold standards (GSs) with perfect sensitivity and specificity as references are identified. Third, these DB diagnoses are subsequently compared with the GS, and the quantitative measurement of validity, including sensitivity, specificity, and positive predictive value will be calculated. Although DB diagnosis definition deemed best is sometimes selected based on the quantitative measurement of validity[2,4], outcome misclassification can cause bias on the estimation unless both the sensitivity and specificity of the outcome definition equal 1.0. Indeed, when the prevalence of outcome of interest is very low, the error-prone outcome definitions having a specificity which are a bit away from 1.0 cause severe bias on estimation[5]. Thus, some authors have recommended that researchers conduct bias-analysis for outcome misclassification based on the performance of outcome definition rather than just using the outcome definition with preferable performance[6].

Although validation studies have been conducted to assess the accuracy of disease diagnosis[4], many have overlooked the possibility of misclassification in medical charts and disease registries serving as GS references. The misclassification of reference data can introduce bias into the findings of epidemiological studies based on DB data. While some researchers have acknowledged the potential for misclassification in GS reference when investigating the impact of residual misclassification in hypothetical studies[7,8], a method to quantitatively evaluate the extent of possible bias and loss of validity in validation studies has not been fully explored. Furthermore, researchers often assume independence between DB diagnoses, meaning that a positive (or negative) result from one diagnostic definition is not associated with a positive (or negative) result from another diagnostic definition. However, it is natural to expect interdependence among the DB diagnoses derived from the same data source.

To address these issues, we propose a novel approach for quantitatively assessing the bias in validation studies. This approach utilizes the latent class model (LCM), which is a statistical tool used to model a class variable with an unknown true status. Since the development of the Hui and Walter model[9] as the foundation of LCMs, various LCMs have been employed to adjust the biases stemming from imperfect diagnostic tests and dependence among DB diagnoses by incorporating a conditional dependence structure as a fixed effect[10–12] or random effect[13,14], explanatory covariates[15], non-constant accuracy rates[16], and extension to the Bayesian approach[17]. However, the application of LCMs in validation studies remains underexplored.

Our research introduces a new application of Bayesian latent class model (BLCM)[17] to validation studies. Although implementing BLCM can be challenging owing to the privacy protection of subject-level datasets, there are scenarios in which the frequency data can be replicated from summary tables. Specifically, we adopted BLCM[17] for a case involving three diagnostic tests, where the frequency data could be reproduced from the summary table of Sato et al.[18]. The BLCM allows for the consideration of constraints on specific parameters, aligning with the conventional setting of validation studies, including the GS assumption. Therefore, the reproducibility of the existing study results can be assessed and compared with the results of another setting.

In this study, we evaluated various models, ranging from those with strict assumptions akin to existing validation studies, such as GS and conditional independence, to those in which these assumptions were relaxed. Our results indicate that when neither GS nor conditional independence was assumed, the indices of diagnostic accuracy deviated significantly from those of existing validation studies and outperformed other models. These findings suggest that existing outcome validation assessments may introduce bias into DB diagnoses and warrant reevaluation using the proposed approach.

METHODS

NOTATION AND MODEL

We now explain the notation used in this study. First, we represent the results of the diagnostic tests represented by T_{jk}, where k(=1, ..., K) corresponds to subject k and j(=1, ..., J) corresponds to the test, with $T_{jk} = 1$ indicating a positive result and $T_{jk} = 0$ a negative result.

A binary latent variable, D_k, represents the presence or absence of the disease of interest in subject k, with $D_k = 1$ indicating the presence of the disease and $D_k = 0$ indicating its absence. Since a disease cannot be directly observed and can only be inferred from the results of imperfect tests, we define the prevalence of the disease in a population as $\mathrm{P}(D = 1)$. To note, we omit the subject's index k because fixed-effects models are considered. The accuracy indices for diagnostic tests are formulated as follows. The sensitivity of test j is considered a random variable, denoted as S_j, and is defined as the conditional probability of a positive test result given that those subjects have a disease ($S_j = \mathrm{P}(T_j = 1|D = 1)$). The specificity of test j is also treated as a random variable and denoted as C_j(= $\mathrm{P}(T_j = 1|D = 0)$). To account for situations where the true disease status cannot be observed, we introduce a latent parameter π representing the prevalence of the disease in a population ($\pi = \mathrm{P}(D = 1)$). With the prevalence, sensitivity and specificity, positive predictive value (PPV) is expressed as $\mathrm{P}(D = 1|T_j = 1) = \pi S_j/(\pi S_j + (1 - \pi)(1 - C_j))$. Negative predictive value (NPV) is also represented as $\mathrm{P}(D = 0|T_j = 0) = (1 - \pi)C_j/(\pi(1 - S_j) + (1 - \pi)C_j)$. The unconditional probability of a diagnostic result can then be expressed as the sum of two conditional probabilities, given the latent class D(=0, 1), as follows:

$$\begin{aligned}&\mathrm{P}(T_j = 1)\\&\quad = \pi\mathrm{P}(T_j = 1 \mid D = 1) + (1 - \pi)\mathrm{P}(T_j = 1 \mid D = 0)\end{aligned} \quad (1)$$

In addition, we model the conditional dependence between pairs of tests (indices with l(=1, ..., J) and h(>l, =2, ..., J) by introducing covariances for both sensitivity ($\mathrm{Cov}(S_l, S_h)$) and specificity ($\mathrm{Cov}(C_l, C_h)$).

$$\begin{aligned}&P(T_l = 1, T_h = 1)\\&\quad = \pi\Big\{\mathrm{P}(T_l = 1 \mid D = 1)\mathrm{P}(T_h = 1 \mid D = 1)\\&\qquad + (-1)^{|t_l - t_h|}\mathrm{Cov}(S_l, S_h)\Big\}\\&\qquad + (1 - \pi)\Big\{\mathrm{P}(T_l = 1 \mid D = 0)\mathrm{P}(T_h = 1 \mid D = 0)\\&\qquad + (-1)^{|t_l - t_h|}\mathrm{Cov}(C_l, C_h)\Big\}\end{aligned} \quad (2)$$

Note that t_l, t_h are the observed test results for subject k and tests l and h(>l). Based on this setting, we derive the likelihood function (see the **Supplemental Document**). Formulas for the likelihood function and conditional distributions for general cases can be found in[17].

MOTIVATING EXAMPLE AND SETTING

In this study, we reinvestigate the data from Sato et al.[18] as a motivating example. Their study assessed the accuracy of definitions for identifying breast cancer cases using 14 distinct definitions derived from medical claims data primarily sourced from Diagnosis Procedure Combination (DPC) claims[19]. These claims include information on disease diagnosis codes, surgical procedures, laboratory tests, drug use, and radiation therapy. Furthermore, they identified 633 breast cancer cases as the reference GS subjects by referencing the in-house cancer registry at St. Luke's International Hospital. Using this GS, they estimated the sensitivity, specificity, and positive predictive value (PPV) for these 14 definitions, which were aggregated within the DPC claims for 50,056 participants in 2011.

Although subject-level datasets, including motivating examples, are generally unavailable in existing validation studies, it is possible to replicate frequency data from summary tables for specific patterns. For example, the presence of a positive result under one definition is a necessary condition for a positive result under another. By establishing the necessary conditions as constraints among multiple definitions, it is possible to reproduce the frequencies from a summary table. In this study, among the 14 diagnoses examined by Sato et al.[18], we focused on two specific diagnostic definitions: Definition 1, the broadest definition that merely mentions a single condition "with a diagnosis" (it corresponds to Definition 1 in Sato et al.[18]), and Definition 2, which includes an additional condition "with surgery, chemotherapy, drug therapy, or radiation therapy" (it corresponds to Definition 12 in Sato et al.[18]). Notably, Sato et al.[18] found that Definition 2 demonstrated the best performance among all definitions. Additionally, we consider the registry diagnosis as the reference, which Sato et al.[18] regarded as the GS, and then total $J = 3$ (j = 1, 2[DBs], 3[registry]) in this study.

Both the DB and registry diagnoses were categorized as dichotomous values: + (positive) or – (negative). The corresponding frequencies of the results for each DB diagnosis (1 and 2) and registry diagnosis are summarized in **Table 1**. The corresponding probabilities are summarized in a 4 × 2 contingency table (**Table 2**). Notably, all components in the third row are zero because a positive result in Definition 1 is a prerequisite for a positive result in Definition 2.

Table 1 Frequencies corresponding to the results of DB diagnoses 1 and 2 and registry diagnosis

DB1[a]	DB2[b]	Registry[c]	Frequency
+	+	+	572
+	+	−	83
+	−	+	53
+	−	−	242
−	+	+	0
−	+	−	0
−	−	+	8
−	−	−	49,098

[a] DB diagnosis is defined as "with a diagnosis" only.
[b] DB diagnosis is defined as "with a diagnosis" and "with surgery, chemotherapy, drug therapy, or radiation therapy."
[c] Registry diagnosis as a reference.
+, positive; −, negative, as diagnosis results.

Table 2 Contingency table associated with the results of the definitions of diagnoses

		Registry diagnosis[a]		Marginal probability
		(+)	(−)	
DB diagnoses (DB1, DB2)[b]	(+, +)	$P(T_1 = 1, T_2 = 1, T_3 = 1)$	$P(T_1 = 1, T_2 = 1, T_3 = 0)$	$P(T_1 = 1, T_2 = 1)$
	(+, −)	$P(T_1 = 1, T_2 = 0, T_3 = 1)$	$P(T_1 = 1, T_2 = 0, T_3 = 0)$	$P(T_1 = 1, T_2 = 0)$
	(−, +)	0	0	0
	(−, −)	$P(T_1 = 0, T_2 = 0, T_3 = 1)$	$P(T_1 = 0, T_2 = 0, T_3 = 0)$	$P(T_1 = 0, T_2 = 0)$
Marginal probability		$P(T_3 = 1)$	$P(T_3 = 0)$	1

[a] Registry diagnosis as a reference.
[b] In definition 1 (DB1) considers only a single condition such as a disease code. In Definition 2 (DB2), additional conditions such as treatment and procedure were considered.
+, positive; −, negative, as results for each diagnostic definition.

Although this study presents a formulation that includes only two definitions in terms of data availability, it can be readily extended to more general scenarios (i.e., involving four or more diagnostic definitions) once subject-level datasets become available.

SETTING FOR BAYESIAN INFERENCE

To explore various scenarios, we investigate four models based on assumptions regarding GS for registry diagnosis ($S_3 = C_3 = 1$) and conditional independence ($\mathrm{Cov}(S_l, S_{h(>l)}) = \mathrm{Cov}(C_l, C_{h(>l)}) = 0$). Model 1 assumes both GS and conditional independence. Model 2 assumes conditional independence but not GS. Model 3 assumes GS but not conditional independence. Finally, Model 4 assumes neither GS nor conditional independence.

The number of parameters to be estimated varies depending on the model. Specifically, in the most complex Model 4, the number of parameters is 13 (prevalence, three sensitivities, three specificities, and six covariances), which exceeds the degrees of freedom of the model represented by the number of independent multinomial cell frequencies for three diagnostic tests minus one ($7 = 2^3 - 1$). Therefore, we opt for Bayesian estimation. In Bayesian inference, the prior distributions are defined as follows. First, the prevalence (π) and sensitivity and specificity are assumed to follow beta distributions with distinct parameters:

$$\pi \sim \mathrm{beta}(\alpha_\pi, \beta_\pi) \tag{3}$$

$$S_j \sim \text{beta}\left(\alpha_{S_j}, \beta_{S_j}\right) \quad (4)$$

$$C_j \sim \text{beta}\left(\alpha_{C_j}, \beta_{C_j}\right) \quad (5)$$

The covariances of sensitivity and specificity are also assumed to follow beta distributions with boundary constraints. The upper bound is included to maintain the sensitivity and specificity within a range of 0 to 1. The lower bound is included because the two diagnostic tests are expected to positively correlate.

$$0 \le \text{Cov}\left(S_l,\ S_{h(>l)}\right) \le \min\left(S_l,\ S_h\right) - S_l S_h \quad (6)$$

$$0 \le \text{Cov}\left(C_l,\ C_{h(>l)}\right) \le \min\left(C_l,\ C_h\right) - C_l C_h \quad (7)$$

$$\text{Cov}\left(S_l,\ S_{h(>l)}\right) \sim \text{beta}\left(\alpha_{\text{cov}s_{l,h}}, \beta_{\text{cov}s_{l,h}}\right) \quad (8)$$

$$\text{Cov}\left(C_l,\ C_{h(>l)}\right) \sim \text{beta}\left(\alpha_{\text{cov}c_{l,h}}, \beta_{\text{cov}c_{l,h}}\right) \quad (9)$$

We assume that the frequencies corresponding to each possible combination of diagnostic test results (n_{t_1,t_2,t_3}) follow a multinomial distribution with probabilities p_{t_1,t_2,t_3}, and a total of K subjects.

$$n_{t_1,t_2,t_3} \sim \text{Multinomial}\left(p_{t_1,t_2,t_3},\ K\right) \quad (10)$$

The probabilities p_{t_1,t_2,t_3}, which is a simplified form of $\text{P}(T_1 = t_1,\ T_2 = t_2,\ T_3 = t_3)$, are defined from prevalence (π), sensitivities (S_j), specificities (C_j), and these covariances, e.g.,

$p_{1,1,1} = \pi(S_1S_2S_3 + \text{Cov}(S_1, S_2) + \text{Cov}(S_1, S_3) + \text{Cov}(S_2, S_3) + (1 - \pi)((1 - C_1)(1 - C_2)(1 - C_3) + \text{Cov}(C_1, C_2) + \text{Cov}(C_1, C_3) + \text{Cov}(C_2, C_3))$.

In addition to estimating the posterior distribution of each parameter, we calculate the widely applicable information criterion (WAIC)[20] for model comparison. This approach is chosen over a deviance-based criterion because the latent class model is a singular statistical model.

For each parameter, to eliminate arbitrariness, we assume non-informative priors, with $\alpha_\pi = \beta_\pi = 1$ for disease prevalence, $\alpha_{S_j} = 1, \beta_{S_j} = 1$ for sensitivity, and $\alpha_{C_j} = 1, \beta_{C_j} = 1$ for specificity, and $\alpha_{\text{Cov}(S_l,\ S_h)} = \beta_{\text{Cov}(S_l,\ S_h)} = \alpha_{\text{Cov}(C_l,\ C_h)} = \beta_{\text{Cov}(C_l,\ C_h)} = 1$ for covariances. However, given the purpose of diagnostic testing and the expectation of low breast cancer prevalence, high specificity a high NPV of >99%, respectively, and a sensitivity of >50%, informative priors may be considered for better convergence, especially considering non-identifiability[21].

In the specified scenario, we run the Gibbs sampler with iterations ranging from 50,000 to 500,000 and a thinning interval of 10 to 500, depending on the model's complexity, after discarding the initial 10,000/50,000 iterations as adaptation/burn-in. Three chains are employed, each initialized close to the expected results.

SIMULATION SETTING

The aim of the simulations is to assess the accuracy of the four models proposed in the previous section when applied to simulated datasets generated for various scenarios. The process of simulating frequencies for the results of three diagnostic tests (shown in **Table 3**, note that it is rounded to three decimal places) was as follows. As an initial step, we set probabilities following a multinomial distribution based on the estimated results of each model. For instance, in line with Model 4, we set the parameter value π as 0.016127, (S_1, S_2, S_3) as (0.92468, 0.75175, 0.72003), (C_1, C_2, C_3) as (0.99572, 0.99878, 0.99872), and these covariances. These parameter values allowed us to define the expected probabilities of the multinomial distribution, such as $p_{1,1,1}$, which were expected to be 0.011411 (see Equation (10) and the following definitions). Detailed lists of the parameter settings and expected probabilities are provided in **Supplemental Table 1**.

We then generated frequency data for each model scenario, comprising 10,000 participants from a multinomial distribution. This random sampling process was repeated 1,000 times, resulting in 4,000 datasets. For these 4,000 datasets, we applied Bayesian estimation using Models 1–4; these settings were similar to those used in the primary analyses. In addition, a frequentist model (corresponding to Sato et al.[18] approach) was also applied as follows; prevalence, sensitivities, and specificities were estimated from simulated frequencies as $n_{\cdot,\cdot,\cdot,1}/N$, $S_1 = n_{1,\cdot,\cdot,1}/n_{\cdot,\cdot,\cdot,1}$, $S_2 = n_{\cdot,1,1,1}/n_{\cdot,\cdot,\cdot,1}$, $C_1 = n_{0,\cdot,\cdot,0}/n_{\cdot,\cdot,\cdot,0}$, $C_2 = n_{\cdot,0,0,0}/n_{\cdot,\cdot,\cdot,0}$.

As a measure of bias for prevalence, sensitivities and specificities were investigated as the difference between estimated values and the true values set in the simulation as $\left(\widehat{\pi} - \pi\right)$, $\left(\widehat{S_j} - S_j\right)$, $\left(\widehat{C_j} - C_j\right)$ for each simulation step and then calculated those means and standard deviations. For a comparison of predictivity, we calculated the WAIC[20] (except for the frequentist's model). Additionally, we calculated a Pearson-type statistic as a measure of goodness-of-fit[22], defined from the observed frequencies generated during the simulation (n_{t_1,t_2,t_3}) and expected frequencies estimated from each model ($\widehat{n_{t_1,t_2,t_3}} = \widehat{p_{t_1,t_2,t_3}} \times K$) for various test combinations (t_1, t_2, $t_3 = 0, 1$) as follow:

$$G = \sum_{t_1, t_2, t_3 = 0,1} \frac{\left(n_{t_1, t_2, t_3} - \widehat{n_{t_1, t_2, t_3}}\right)^2}{\widehat{n_{t_1, t_2, t_3}}} \quad (11)$$

RESULTS

ANALYSIS RESULTS

Table 3 shows the posterior means of the parameters for all model patterns, along with the reference results of Sato et al.[18]. The key findings are as follows. In Model 1, which closely mirrors the setting of Sato et al.[18] model, our results aligned closely with theirs. Disease prevalence was estimated to be 0.013, and the sensitivity and PPV of DB Definition 2 (0.902, 0.872) were preferable to those of DB Definition 1 (0.986, 0.657). In Model 2, in which the registry diagnosis did not consider GS, the results differed slightly from those of Model 1. The disease prevalence was slightly higher at 0.014 and the sensitivity of the registry diagnosis was estimated at 0.872, approximately ten percentage points lower than that of Model 1. A similar trend was observed for diagnostic Definitions 1 and 2 of the DB diagnoses, with slight increases in sensitivity and PPV (0.998 and 0.753 for Definition 1 and 0.914 and 0.998 for Definition 2, respectively). Model 4, which assumes conditional dependence and lacks GS, yielded distinct results. Disease prevalence was slightly higher (0.016), and sensitivities were estimated to be substantially lower in the other three models: 0.925 in Definition 1 and 0.752 in Definition 2. However, the PPVs were slightly higher (0.779 for Definition 1 and 0.908 for Definition 2). Specificities and NPVs were consistently high (> 99 %) for all diagnostic definitions across model

Table 3 Results summary table: Posterior means of the prevalence and test parameters obtained from four different models with Sato et al.'s[18] results as a reference.

		Conditional independence		Conditional dependence	
		GS	Non-GS	GS	Non-GS
	Sato (2015)	Model 1	Model 2	Model 3	Model 4
π	0.013	0.013	0.014	0.014	0.016
S_1	0.987	0.986	0.998	0.928	0.925
S_2	0.904	0.902	0.914	0.855	0.752
S_3	—	—	0.872	—	0.720
C_1	0.993	0.993	0.995	0.994	0.996
C_2	0.998	0.998	1.000	0.999	0.999
C_3	—	—	1.000	—	0.999
PPV_1	0.658	0.657	0.753	0.689	0.779
PPV_2	0.873	0.872	0.998	0.933	0.908
PPV_3	—	—	0.986	—	0.901
NPV_1	1.000	1.000	1.000	0.999	0.999
NPV_2	0.999	0.999	0.999	0.998	0.996
NPV_3	—	—	0.998	—	0.995
WAIC	—	12,454.0	11,536.8	11,618.6	11,536.5

GS, a model in which registry diagnosis was assumed to be the gold standard; non-GS, a model in which registry diagnosis was assumed to be the gold standard.
π, prevalence; S_j, sensitivity; C_j, specificity; PPV_j, positive predictive value; NPV_j, negative predictive value; subscript j is an index of diagnostic definitions where j = 1, 2 [DBs], 3 [registry]. WAIC, a widely applicable information criterion, is used for model comparison.
For each parameter, to eliminate arbitrariness, we assume non-informative priors, with $\alpha_\pi = \beta_\pi = 1$ for disease prevalence, $\alpha_{S_j} = 1, \beta_{S_j} = 1$ for sensitivity, and $\alpha_{C_j} = 1, \beta_{C_j} = 1$ for specificity, and $\alpha_{\mathrm{Cov}(S_l, S_h)} = \beta_{\mathrm{Cov}(S_l, S_h)} = \alpha_{\mathrm{Cov}(C_l, C_h)} = \beta_{\mathrm{Cov}(C_l, C_h)} = 1$ for covariances.
The Gibbs sampler has iterations ranging from 50,000 to 500,000 and a thinning interval of 10 to 500, depending on the model's complexity, after discarding the initial 10,000 and 50,000 iterations as adaptation and burn-in, respectively.

patterns. Model 4 exhibited the lowest WAIC value, demonstrating the best predictive performance among all the models.

Furthermore, when considering alternative combinations of diagnostic definitions for DB "with a diagnosis" AND "with a diagnosis code related to breast cancer or marker test code," similar results were observed (frequencies are detailed in **Supplemental Table 2**, and results are shown in **Supplemental Table 3**).

SIMULATION RESULTS

Table 4 presents means and standard deviations of bias for prevalence, sensitivities, and specificities. In summary, the following trends were confirmed in all data generation Scenarios 1 to 4. Regarding the bias in prevalence, the Bayesian model corresponding to each scenario (results on the diagonal elements) showed the minimum values, or in other cases, a value that was almost the same as the minimum values. Regarding sensitivities, there was a tendency for the model corresponding to each scenario (results on the diagonal elements) to correspond to the smallest values or the second smallest values. Model 4 differed from the other models in that it showed a substantial bias in Scenarios 1 to 3, with the absolute value reaching a maximum of about 0.3 in some cases. Regarding specificity, in all scenarios, the bias values for each model remained at around three decimal places, indicating minimal bias. In the frequentist's model, it showed better results in Scenarios 1 to 3, with the mean bias value being the same or smaller than any of the Bayesian models. However, in some cases, the estimates fell into corner solutions (probability values of 0 or 1). In Scenarios 1, 2, and 4, such cases were observed in about 10–15% of the total number of simulations.

Table 5 summarizes the median Pearson-type statistics and WAIC for each data-generating scenario and model. Model 4 consistently achieved the lowest Pearson-type statistics and WAIC across all scenarios, except for the WAIC in Scenario 1. These results demonstrate the consistently favorable performance of Model 4, regardless of the data-generating scenario. Notably, the diagonal elements in the table do not significantly differ from the lowest values, suggesting that each model is suitable for its respective data-generating scenarios. The frequentist model had smaller goodness-of-fit values than any of the Bayesian models in Scenario 1. However, in the other scenarios, it showed substantially larger values than any other Bayesian model. Additionally, as previously mentioned, it should be noted that the goodness-of-fit index could not be calculated in some cases owing to the corner solutions for probabilities.

DISCUSSION

This study utilized a Bayesian latent class model to investigate the diagnostic accuracy of publicly available data. Our analysis revealed discrepancies from the existing findings when we assumed conditional dependence and a non-GS diagnosis. Disease prevalence was slightly higher (0.016), with lower sensitivities estimated at 0.925 in Definition 1 and 0.752 in Definition 2. Conversely, slightly higher PPVs were observed at 0.779 for Definition 1 and 0.908 for Definition 2. Model 4 exhibited the best predictive performance among all investigated models.

In the simulation study, the results of the bias evaluation showed that the Bayesian models with more assumptions and the frequentist's model performed better under the true value conditions. However, although the estimates of the diagnostic tests from the frequentist's model were close to the true values, unfavorable results were obtained regarding goodness-of-fit. A possible reason for this is, as mentioned above, the probability of diagnostic tests falling into a corner solution (goodness-of-fit cannot be defined) in some cases. Furthermore, conditional independence and GS assumptions may be corrupted in the calculation of multinomial probabilities based on estimates of diagnostic accuracy. The Bayesian model with fewer assumptions performed well in terms of goodness-of-fit statistics and WAIC, regardless of the data-generating scenario. Because the true value is generally unknown, we believe that it is appropriate to primarily use the Bayesian model with fewer constraints, compare other models, and determine the most favorable diagnostic definition.

These findings are of paramount significance because the existing validation study results have been widely utilized to determine outcome definitions in subsequent epidemiological studies. A biased outcome definition may introduce bias in subsequent epidemiological studies that rely on these definitions. Although some epidemiological studies have incorporated sensitivity analyses to account for the uncertainty in outcomes, accurately quantifying bias has been challenging. The importance of sensitivity analyses remains unquestioned, and the Bayesian latent class model may help in quantitative bias assessment.

This study had some limitations. First, it is a single-case evaluation based on a limited number of diagnostic definitions from a single case study owing to data limitations. Therefore, the generalizability of these findings

Table 4 Summary of results: Means and standard deviations (SDs) of bias for prevalence, sensitivity, and specificity

Mean (SD)		Conditional independence			Conditional dependence	
		Sato (2015)[a]	GS Model 1	Non-GS Model 2	GS Model 3	Non-GS Model 4
Scenario 1 (CI & GS)	π	0.000 (0.001)	0.000 (0.001)	0.000 (0.001)	0.000 (0.001)	0.003 (0.001)
	S_1	−0.000 (0.010	−0.019 (0.010)	−0.018 (0.009)	−0.024 (0.009)	−0.140 (0.015)
	S_2	−0.001 (0.027)	−0.013 (0.025)	−0.014 (0.025)	−0.029 (0.025)	−0.209 (0.027)
	S_3	—	—	−0.024 (0.003)	—	−0.297 (0.008)
	C_1	0.000 (0.001)	0.000 (0.001)	0.000 (0.001)	−0.001 (0.001)	0.000 (0.001)
	C_2	0.000 (0.000)	0.005 (0.000)	0.005 (0.000)	0.005 (0.000)	0.004 (0.000)
	C_3	—	—	0.006 (0.000)	—	0.004 (0.000)
Scenario 2 (CI & non-GS)	π	−0.002 (0.001)	−0.002 (0.001)	0.000 (0.001)	−0.001 (0.001)	0.003 (0.001)
	S_1	−0.014 (0.011)	−0.033 (0.010)	−0.018 (0.003)	−0.092 (0.016)	−0.100 (0.006)
	S_2	−0.013 (0.026)	−0.025 (0.024)	−0.014 (0.023)	−0.079 (0.024)	−0.228 (0.018)
	S_3	—	—	−0.010 (0.027)	—	−0.213 (0.022)
	C_1	−0.002 (0.001)	−0.002 (0.001)	0.000 (0.001)	−0.001 (0.001)	0.000 (0.001)
	C_2	−0.002 (0.000)	0.003 (0.000)	0.004 (0.000)	0.004 (0.000)	0.003 (0.000)
	C_3	—	—	0.004 (0.000)	—	0.003 (0.000)
Scenario 3 (CD & GS)	π	0.000 (0.001)	0.000 (0.001)	0.000 (0.001)	0.000 (0.001)	0.003 (0.001)
	S_1	0.000 (0.022)	−0.013 (0.021)	0.050 (0.003)	−0.023 (0.017)	−0.057 (0.011)
	S_2	−0.000 (0.030)	−0.008 (0.028)	0.048 (0.023)	−0.021 (0.026)	−0.182 (0.017)
	S_2	—	—	−0.082 (0.021)	—	−0.284 (0.020)
	C_1	0.000 (0.001)	0.000 (0.001)	0.001 (0.001)	0.000 (0.001)	0.001 (0.001)
	C_2	0.000 (0.000)	0.005 (0.000)	0.005 (0.000)	0.005 (0.000)	0.004 (0.000)
	C_3	—	—	0.004 (0.000)	—	0.003 (0.000)
Scenario 4 (CD & non−GS)	π	−0.003 (0.001)	−0.003 (0.001)	−0.002 (0.001)	−0.002 (0.001)	0.001 (0.001)
	S_1	0.060 (0.011)	0.041 (0.010)	0.055 (0.003)	−0.019 (0.016)	−0.026 (0.006)
	S_2	0.149 (0.027)	0.137 (0.025)	0.148 (0.024)	0.082 (0.024)	−0.066 (0.018)
	S_3	—	—	0.141 (0.027)	—	−0.061 (0.022)
	C_1	−0.002 (0.001)	−0.003 (0.001)	−0.001 (0.001)	−0.002 (0.001)	0.000 (0.001)
	C_2	−0.000 (0.000)	0.002 (0.000)	0.004 (0.000)	0.003 (0.000)	0.002 (0.000)
	C_3	—	—	0.004 (0.000)	—	0.002 (0.000)

[a] This model is associated with frequentist's model: prevalence, sensitivities, and specificities were estimated from simulated frequencies as $n_{\ldots,1}/N$, $S_1 = n_{1,\ldots,1}/n_{\ldots,1}$, $S_2 = n_{\cdot,1,1}/n_{\ldots,1}$, $C_1 = n_{0,\ldots,0}/n_{\ldots,0}$, $C_2 = n_{\cdot,0,0}/n_{\ldots,0}$. In this model, some results were corner solutions (probability values of 0 or 1); in these cases, the goodness of fit could not be calculated. In Scenarios 1, 2, and 4, such cases were observed in approximately 10%–15% of the total number of simulations.

GS, a model in which registry diagnosis was assumed to be the gold standard; non-GS, a model in which registry diagnosis was assumed to not be the gold standard.

π, prevalence; S_j, sensitivity; C_j, specificity; PPV_j, positive predictive value; NPV_j, negative predictive value; subscript j is an index of diagnostic definitions where j = 1, 2 [DBs], 3 [registry]. CI, conditional independence; CD, conditional dependence.

The shaded numbers show the minimum values of the mean bias among the Bayesian models for each parameter in each scenario. Note that the minimum values of the mean bias were not identified for S_3 and C_3 in Scenarios 1 and 3 (assumed GS) because these values were set deterministically to one.

Generated frequency data for each model scenario comprising 10,000 subjects from a multinomial distribution. This random sampling process was repeated 1,000 times, resulting in 4,000 datasets. For these 4,000 datasets, we applied Bayesian estimation using Models 1 to 4, with settings similar to those used in the primary analyses, as well as the frequentist's model.

Table 5 Median Pearson type statistics and WAIC for the data-generating scenarios and models

Upper: median Pearson-type statistics[a] Lower: median WAIC		Conditional independence		Conditional dependence	
		GS	Non-GS	GS	Non-GS
	Sato (2015)[b]	Model 1	Model 2	Model 3	Model 4
Scenario 1	0.4 NA	2.9 2,503	9.0 2,515	5.6 2,510	2.7 2,505
Scenario 2	2,486.0 NA	1,957.0 2,509	9.3 2,341	15.8 2,349	3.1 2,330
Scenario 3	1,430.0 NA	941.1 2,457	8.2 2,345	3.4 2,336	2.6 2,336
Scenario 4	2,485.0 NA	1,954.0 2,508	9.3 2,339	15.7 2,345	3.1 2,328

[a] The Pearson type statistics defined as the square of (observed frequencies generated from simulation—expected frequencies estimated from the Bayesian model) divided by expected frequencies.
[b] This model is associated with frequentist's model: prevalence, sensitivities, and specificities were estimated from simulated frequencies as $n_{\cdot\cdot\cdot 1}/N$, $S_1 = n_{1\cdot\cdot 1}/n_{\cdot\cdot\cdot 1}$, $S_2 = n_{\cdot 1\cdot 1}/n_{\cdot\cdot\cdot 1}$, $C_1 = n_{0\cdot\cdot 0}/n_{\cdot\cdot\cdot 0}$, $C_2 = n_{\cdot 0\cdot 0}/n_{\cdot\cdot\cdot 0}$. In this model, some results were corner solutions (probability values of 0 or 1); in these cases, the goodness of fit could not be calculated. In scenarios 1, 2, and 4, such cases were observed in approximately 10%–15% of the total number of simulations.
In each scenario 1–4, the observed frequency data of 10,000 subjects were generated 1,000 times via random sampling from a multinomial distribution with these parameters set similar to each model 1–4.
The shaded numbers show the minimum values of the median Pearson type statistics and median WAIC in each scenario.
WAIC, a widely applicable information criterion, is used for model comparison.

should be assessed using data from other studies. Second, this study employs classical methods, and there are more sophisticated approaches, such as those involving higher-order correlation parameters[23] and hierarchical structures[24]. Future studies on the implementation of these advanced methods are required. Further to note, we made two key assumptions: non-differential misclassification of the outcome[25] and consistent performance of the outcome definition across the entire population in the validation study. If there are concerns about differential misclassification among factors or the consistency of effects across the population, it may be feasible to extend the model. This could involve incorporating exposure factors or subject-specific information as explanatory variables in the latent class model to evaluate their impact. In such scenarios, however, it is crucial to pay heightened attention to the convergence of the Markov Chain Monte Carlo due to the increased complexity of a model.

In our study, we advocate for the use of non-informative priors to mitigate researcher bias and ensure the robustness of Markov Chain Monte Carlo convergence. However, we acknowledge that in validation studies, prior knowledge on disease prevalence and diagnostic test accuracy can be beneficial. As such, incorporating this preliminary information is deemed acceptable. This approach aligns with the approach applied by Pereira da Silva et al.[21], which involves eliciting expert opinions and integrating the derived minimum and maximum values into the parameters of the beta distribution. This method effectively leverages existing knowledge while maintaining scientific rigor.

Moreover, in the context of validation studies, the selection of the most appropriate diagnostic definition can be a formidable challenge, particularly when no single definition uniformly outperforms the others. Even when a superior definition is available, a consensus on the optimal choice may not be attained. It is important to emphasize that the assessment undertaken in this study is inherently model based and founded on statistical evaluations, representing only one facet of the broader process. A comprehensive assessment encompassing the clinical validity and other relevant considerations should guide the decision. When implementing diagnostic definitions derived from validation studies in practical epidemiological research, keen awareness of the potential for outcome definitions to deviate from the true underlying states is required, and a diverse sensitivity analysis should be considered. While this study primarily offers a statistical evaluation within the scope of outcome validation, it underscores the significance of holistic decision-making, thoughtful sensitivity analyses, and informed interpretation of results in the broader landscape of epidemiology.

CONCLUSIONS

In conclusion, our study emphasizes that the practice of assuming registry diagnosis as the GS can introduce bias into DB diagnosis. Thus, the proposed approach is a valuable tool for consideration in validation studies. In the future, we intend to conduct further research to explore more advanced methods and apply them to actual case data.

CONFLICTS OF INTEREST STATEMENT

Satoshi Uno is an employee of Astellas Pharma Inc. Toshiro Tango has no financial or nonfinancial interests to disclose.

ACKNOWLEDGMENT

The authors wish to thank Hisashi Noma for very kindly reviewing this draft. The authors also wish to thank Editage for English language editing.

REFERENCES

1. Sherman RE, Anderson SA, Dal Pan GJ, et al. Real-World Evidence—what is it and what can it tell us? *N Engl J Med*. 2016;375:2293–7.
2. Ritchey ME, West SL, Maldonado G. Validity of drug and diagnosis data in pharmacoepidemiology. In: Strom BL, Kimmel SE, Hennessy S, editors. *Pharmacoepidemiology*. 2019:948–90.
3. Chun DS, Lund JL, Stürmer T. Pharmacoepidemiology and Drug Safety's special issue on validation studies. *Pharmacoepidemiol Drug Saf*. 2019;28:123–5.
4. Koram N, Delgado M, Stark JH, et al. Validation studies of claims data in the Asia-Pacific region: A comprehensive review. *Pharmacoepidemiol Drug Saf*. 2019;28:156–70.
5. Chubak J, Pocobelli G, Weiss NS. Tradeoffs between accuracy measures for electronic health care data algorithms. *J Clin Epidemiol*. 2012;65:343–9.
6. Lash TL. Bias analysis. In: Lash TL, VanderWeele TJ, Haneuse S, Rothman KJ, editors. *Modern Epidemiology.* 2021:711–54.
7. Setoguchi S, Solomon DH, Glynn RJ, et al. Agreement of diagnosis and its date for hematologic malignancies and solid tumors between medicare claims and cancer registry data. *Cancer Causes Control*. 2007;18:561–9.
8. Li Q, Glynn RJ, Dreyer NA, et al. Validity of claims-based definitions of left ventricular systolic dysfunction in Medicare patients. *Pharmacoepidemiol Drug Saf*. 2011;20:700–8.
9. Hui SL, Walter SD. Estimating the error rates of diagnostic tests. *Biometrics*. 1980;36:167–71.
10. Vacek PM. The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics*. 1985;41:959–68.
11. Sinclair MD, Gastwirth JL. On procedures for evaluating the effectiveness of reinterview survey methods: application to labor force data. *J Am Stat Assoc*. 1996;91:961–9.
12. Torrance-Rynard VL, Walter SD. Effects of dependent errors in the assessment of diagnostic test performance. *Stat Med*. 1997;16:2157–75.
13. Qu Y, Tan M, Kutner MH. Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics*. 1996;52:797–810.
14. Hadgu A, Qu Y. A biomedical application of latent class models with random effects. *J R Stat Soc Ser C Appl Stat*. 1998;47:603–16.
15. Frössling J, Bonnett B, Lindberg A, et al. Validation of a Neospora caninum iscom ELISA without a gold standard. *Prev Vet Med*. 2003;57:141–53.
16. Brenner H, Gefeller O. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. *Stat Med*. 1997;16:981–91.
17. Dendukuri N, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics*. 2001;57:158–67.
18. Sato I, Yagata H, Ohashi Y. The accuracy of Japanese claims data in identifying breast cancer cases. *Biol Pharm Bull*. 2015;38:53–7.
19. Wang K, Li P, Chen L, et al. Impact of the Japanese Diagnosis Procedure Combination-based Payment System in Japan. *J Med Syst*. 2010;34:95–100.
20. Watanabe S. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J Mach Learn Res*. 2010;11:3571–94.
21. Pereira da Silva HD, Ascaso C, Goncalves AQ, et al. A Bayesian approach to model the conditional correlation between several diagnostic tests and various replicated subjects measurements. *Stat Med*. 2017;36:3154–70.
22. Schofield MR, Maze MJ, Crump JA, et al. On the robustness of latent class models for diagnostic testing with no gold standard. *Stat Med*. 2021;40:4751–63.
23. Wang Z, Dendukuri N, Zar HJ, et al. Modeling conditional dependence among multiple diagnostic tests. *Stat Med*. 2017;36:4843–59.
24. Wang C, Lin X, Nelson KP. Bayesian hierarchical latent class models for estimating diagnostic accuracy. *Stat Methods Med Res*. 2019;29:1112–28.
25. Gusafson P. The impact of unacknowledged measurement error. In: Yi GY, Delaigle A, Gustafson P, editors. *Handbook of Measurement Error Models*. 2021:37–52.