Published in final edited form as:
Stat Med. 2021 September 10; 40(20): 4430–4441. doi:10.1002/sim.9039.

Statistical insights for crude-rate-based operational measures of misdiagnosis-related harms

Yuxin Zhu[1,2], **Zheyu Wang**[1,2], **Ava L. Liberman**[3], **Tzu-Pu Chang**[4,5], **David Newman-Toker**[6,7,8]

[1]Division of Biostatistics and Bioinformatics, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, Maryland [2]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland [3]Albert Einstein College of Medicine, Montefiore Medical Center, Bronx, New York [4]Department of Neurology/Neuro-Medical Scientific Center, Taichung Tzu Chi Hospital, Buddhist Tzu Chi Medical Foundation, Taichung, Taiwan [5]Department of Neurology, School of Medicine, Tzu Chi University, Hualien, Taiwan [6]Department of Neurology, Johns Hopkins University School of Medicine, Baltimore, Maryland [7]Armstrong Institute Center for Diagnostic Excellence, Baltimore, Maryland [8]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland

Abstract

In longitudinal event data, a crude rate is a simple quantification of the event rate, defined as the number of events during an evaluation window, divided by the at-risk population size at the beginning or mid-time point of that window. The crude rate recently received revitalizing interest from medical researchers who aimed to improve measurement of misdiagnosis-related harms using administrative or billing data by tracking unexpected adverse events following a “benign” diagnosis. The simplicity of these measures makes them attractive for implementation and routine operational monitoring at hospital or health system level. However, relevant statistical inference procedures have not been systematically summarized. Moreover, it is unclear to what extent the temporal changes of the at-risk population size would bias analyses and affect important conclusions concerning misdiagnosis-related harms. In this article, we present statistical inference tools for using crude-rate based harm measures, as well as formulas and simulation results that quantify the deviation of such measures from those based on the more sophisticated Nelson-Aalen estimator. Moreover, we present results for a generalized multibin version of the crude rate, for which the usual crude rate is a single-bin special case. The generalized multibin crude rate is more straightforward to compute than the Nelson-Aalen estimator and can reduce potential biases of the single-bin crude rate. For studies that seek to use multibin measures, we provide simulations to guide the choice regarding number of bins. We further bolster these results using a worked example of stroke after “benign” dizziness from a large data set.

Correspondence Zheyu Wang, 550 N. Broadway, Room 1111-D, Baltimore, MD 21205-2013. wangzy@jhu.edu.

SUPPORTING INFORMATION
Additional supporting information may be found online in the Supporting Information section at the end of this article.

Keywords

crude rate; cumulative hazard; deviation bound; misdiagnosis-related harm

1 | INTRODUCTION

Diagnostic error is one of the most important safety problems in health care today and inflicts the most cost among all medical errors.[1–3] Traditionally, identification of diagnostic-error-related harm has often relied on a labor-intensive chart review process, which is prone to reviewer bias,[4] restricted by poor documentation,[5,6] and is hard to apply on a large scale. To address this urgent need, medical researchers, in the recent years, have attempted to develop an automatic tool that quantifies harm that results from a "benign" (false negative) diagnosis, utilizing the abundant disease occurrence information in existing clinical, billing, administrative claims, or similar electronic medical record (EMR) data sets.[7–12]

This conceptual work is formalized by Liberman and Newman-Toker.[13] The fundamental idea is that for acute diseases such as stroke, unexpected adverse events attributable to a false negative diagnosis would likely occur in a relatively short time window after the initial diagnosis. Also, the risk of adverse events would have been constant had there been no misdiagnosis of the symptoms. Therefore, the existence of excessive short-term adverse event rate beyond the expected baseline adverse event rate for the population without misdiagnosis would indicate the existence of misdiagnosis of symptoms, and the magnitude of excessive risk reflects the magnitude of harm. Because the baseline rate is an unobservable counterfactual quantity, one way that researchers proposed to approximate this quantity was to use the long-term rate, that is, the event rate in a time window temporally far away from the initial diagnosis and therefore are likely to be independent of misdiagnosis. Using the long-term event rate estimated in the population of interest as a surrogate for the counterfactual baseline rate also have a preferred property that it implicitly accounts for the population composition disparity across institutes. This is especially important for diseases such as stroke, whose baseline event rate is greatly elevated in some subgroups such as in elderly. As a result, comparison measures between the short-term and long-term rates, such as crude rate differences and ratios, can be used to quantify misdiagnosis-related harm magnitude.

Mane et al[14] applied this idea to analyze misdiagnosis related stroke among patients that received a "benign" dizziness diagnosis. They proposed an operational measure based on the widely used crude rate of stroke return visits. The crude rate was defined as

$$\text{crude rate in window } I = \frac{\text{number of events in } I}{\text{risk set size at beginning of } I \times \text{length of } I}.$$

There is a rich statistical literature on more sophisticated estimators such as the Nelson-Aalen,[15,16] but crude-rate based estimators remain widely used in public health and medical research and applications, especially in the area of safety and quality measurement. For example, many of the measures endorsed by the National Quality Forum are crude rates constructed with counts as numerators and denominators. The popularity of crude rates can be attributed to several reasons. First, crude rates are highly accessible to policymakers, health care workers, and patients because they can easily understand such measures. Second, steps to compute crude rates are simple and straightforward and can therefore be easily carried out by researchers. This is especially important for quality measures that are often computed without using sophisticated statistical software. The main computational advantage is not due to the reduced computational complexity, which is at the order of n^2 for a Nelson-Aalen estimator and at the order of n for a single-bin crude rate, but due to the reduced steps in the whole pipeline from data to a harm measure. For example, the data warehouse server usually does not have adequate statistical software, due to limited RAM, restricted administrative access for a statistician, the immense amount and complicated structure of the data, as well as other confidentiality concerns. To calculate the N-A estimator in a statistical software where computation is efficient, it would require that the EMR or large claim databases be extracted, cleaned, and transferred to a statistician, who may also needs approval for accessing the extracted data. The time-consuming data extraction and transfer can only be carried out by limited personnel who have access but who may not have enough statistical sophistication to clean the data into analysis-ready format. Further cleaning needs to be done by a statistician, which also takes time. On the other hand, simple data summaries such as counts usually can be directly obtained from the data server. As a result, measures based on crude rates can be calculated directly without extracting, transferring an analytic data set or obtaining high data access level. Simple and accessible calculations, in turn, result in the timely evaluation of institutes' diagnostic performance that can inform interventions and policies that improve quality of care. Third, the crude rates often yield results that are close to and as powerful as the more sophisticated estimator in applications where event rates and censoring rates are low, which is often the case in safety and quality measurement applications.

Although crude rates often work well in applications, practitioners are not properly guided on specific methods to use and do not understand why crude rates are applicable and, more importantly, when they are not and need to be replaced by a different estimator. This work addresses this gap in understanding and provides guidance. To understand the interpretation, bias, and usefulness of the crude rate, we consider a generalized multibin version of the crude rate, with the standard single-bin crude rate as a special case. Through the multibin crude rate, we establish the connection between the crude rate and the cumulative hazard, which leads to the interpretation of the crude rate as a rough estimator of cumulative hazard. We then study the difference between multibin crude rates and the Nelson-Aalen estimator, and we summarized statistical inferential and hypothesis testing methods that are necessary in an analysis using the general multibin crude rates. In addition, we perform simulation studies to provide insights on situations where the single-bin crude rate may have adequately good performance in terms of measuring and detecting harm. When the single-bin crude rate performs not as well, we recommend that researchers consider general multibin crude rates as alternatives. Further guidance on the choice of number of bins is provided through simulation studies with settings that closely resemble real data. Finally, we illustrate with a data analysis how the information in simulation studies can be used to guide the choice of bins. Overall, we showed with simulation studies and the real data analysis that, when the outcome disease and censoring occur at relatively low rates, the harm measure based on the single-bin crude rate has comparably good performance as the more sophisticated Nelson-Aalen estimator.

2 | MULTIBIN CRUDE RATE

Suppose we want to calculate crude rate of event occurrences during some predetermined time interval I of length L. The danger of using the standard calculation, as we argued in the Introduction, is that we might introduce potentially large bias to the estimator in ignoring the decrease of risk set size over time. To understand the effect of a shrinking risk set and mitigate the effect when it is severe, we consider a generalized multibin version of the crude rate described as follows.

Let I_j, j=1, …, J denote bins that divide interval I, let t_j denote the left endpoint for bin I_j, and let L_j denote the length of I_j. Then the crude rate in time bin I_j is

$$\text{crude rate in window } I_j = \frac{\text{number of events in } I_j}{\text{risk set size at } t_j \times L_j}.$$

We then obtain the multibin crude rate with J bins, or simply the J-bin crude rate, by taking average of the naïve crude rates over J bins weighted by lengths. Or equivalently,

$$J\text{-bin crude rate in } I = \left(\sum_{j=1}^{J} L_j\right)^{-1} \sum_{i=1}^{J} \left(\frac{\text{number of events in } I_j}{\text{risk set size at } t_j \times L_j} \times L_j\right) = L^{-1} \sum_{i=1}^{J} \frac{\text{number of events in } I_j}{\text{risk set size at } t_j}.$$

The multibin crude rate is an improved version of the single-bin crude rate in accuracy and adds little computational complexity. It also serves as a bridge between the naïve estimator and the quantity we aim to measure. Specifically, if we take the effectively finest interval partition, that is, we use bins such that t_j, j=1, …, J, are distinct time points of all event occurrences, then we observe exactly one event in each bin and the decrease in risk set is fully accounted for. In this case, the J-bin crude rate is

$$L^{-1} \sum_{j=1}^{J} \frac{1}{\text{risk set size at } t_j},$$

which is exactly the Nelson-Aalen estimator for cumulative hazard, divided by L. If we further assume that the hazard rate is constant in interval I, this special J-bin crude rate is the hazard over I. Therefore the crude rate can be considered as a rough estimate of the average hazard over interval I, and the difference and ratio between short-term and long-term crude rates can be interpreted as the difference and ratio of average hazards in the misdiagnosed population and approximately the correctly diagnosed population.

Despite similar forms, the multibin crude rates are not a special type of the weighted or scaled N-A estimator.[17] In general, the multibin crude rate has systematic bias and does not converge to the population cumulative hazard even when sample size goes to infinity, but the N-A estimator is a consistent estimator of the population cumulative hazard given some predetermined covariate distribution. However, as partition of time interval becomes finer, the bias of multibin crude rate gets smaller and eventually converges to zero. Theoretically, we can always expect improvement with interval partitions when each bin covers at least one event occurrence, refining the risk set considered at the beginning of each bin. We show in Section 3 that the improvement of a *J*-bin crude rate over a single-bin crude rate depends on how event occurrences and censoring are distributed across bins. In addition, we illustrate the performance of the multibin crude rate in relation to the number of bins with simulation studies in Section 4.

3 | THEORETICAL PROPERTIES

In this section, we study properties of the multibin crude rate estimators. We consider, for patients indexed by i=1, …, n, a general survival outcome T_i which is subject to independent censoring C_i. Suppose (T_i,C_i) are independent and identically distributed across i, and that we observe data (Y_i,Δ_i) where $Y_i = \min(T_i,C_i)$ and $\Delta_i = \mathbb{1}(T_i \le C_i)$. All of the following results cover the standard single-bin crude rate as a special case.

3.1 | Bias in multibin crude rate from the Nelson-Aalen estimator

Consider a time window $I = [t_0, t_J)$, over which we calculate the J-bin crude rate using partition at points $t_0 < t_1 < \cdots < t_J$. This J-bin crude rate is formally defined as

$$\widehat{CR}\{t_j\}_{j=0}^{J} = (t_J - t_0)^{-1} \sum_{j=1}^{J} \frac{\sum_{i=1}^{n} \mathbb{1}(t_{j-1} \le Y_i < t_j,\ \Delta_i = 1)}{\sum_{i=1}^{n} \mathbb{1}(Y_i \ge t_{j-1})},$$

and we are interested in its deviation from the Nelson-Aalen estimator. Suppose that we observe events at time $t_j^1 < \ldots < t_j^{m_j}$ over interval $[t_{j-1}, t_j)$ and that we have $t_j^0 \equiv t_{j-1}$. Denote by $\hat{P}_n(A) = n^{-1} \sum_{i=1}^{n} \mathbb{1}(A)$ the empirical probability of some event A occurring. With some algebra, we have

$$\widehat{CR}\{t_j\}_{j=0}^{J} = (t_J - t_0)^{-1} \sum_{j=1}^{J} \sum_{m=1}^{m_j} \frac{\hat{P}\left(t_j^m \le Y_i < t_j^{m+1},\ \Delta_i = 1\right)}{\hat{P}\left(Y_i \ge t_j^m\right)} \times \frac{\hat{P}\left(Y_i \ge t_j^m\right)}{\hat{P}\left(Y_i \ge t_{j-1}\right)},$$

and its difference from the Nelson-Aalen can be written as

$$\begin{aligned}(t_J - t_0)^{-1} &\sum_{j=1}^{K} \sum_{m=1}^{m_j} \left[\frac{1}{\hat{P}\left(Y_i \ge t_j^m\right)} \times \left\{1 - \frac{\hat{P}\left(Y_i \ge t_j^m\right)}{\hat{P}\left(Y_i \ge t_{j-1}\right)}\right\}\right] \\ = &\sum_{j=1}^{K} \sum_{m=1}^{m_j} \frac{1}{\hat{P}\left(Y_i \ge t_j^m\right)} - \sum_{j=1}^{K} \frac{m_j}{\hat{P}\left(Y_i \ge t_{j-1}\right)}.\end{aligned} \tag{1}$$

We observe that the first summation on the right-hand side of the equation above is invariant to the partition of time interval I and that the quantity above is always positive. Then to minimize the bias of a K-bin crude rate for fixed K, we maximize the quantity $\sum_{j=1}^{K} m_j / \hat{P}\left(Y_i \geq t_{j-1}\right)$. With some algebra we can show that when censoring is ignorable and event rate is low, a partition that assigns each time bin with equal number of events gives the smallest bias. Under more general scenarios, a different partition could yield better results. In practice, however, it is often unrealistic to use partitions with irregular time points that are based on event occurrences as it beats the purpose of using the simplistic crude rates. A common practice is to use the partition with equal bin lengths, which could be sub-optimal but actually creates little additional bias when event and censoring rates are low, as implied by (1).

3.2 | Statistical inference for multibin crude rate difference and ratio

In this section, we summarize statistical inference and hypothesis testing methods for harm measures based on multibin crude rates. We consider a short-term window $I_S = \left[t_0^S, t_{J_1}^S\right)$ over which we calculate the J_1-bin crude rate using bin partition at $t_0^S < t_1^S < \ldots < t_{J_1}^S$, and a long-term window $I_L = \left[t_0^L, t_{J_2}^L\right)$ over which we calculate the J_2-bin crude rate using partitions at $t_0^L < t_1^L < \ldots < t_{J_1}^L$. Let $\widehat{CR}\left\{t_j^S\right\}_{j=0}^{J_1}$ and $\widehat{CR}\left\{t_j^L\right\}_{j=0}^{J_2}$ denote these two crude rates, and let $CR\left\{t_j^S\right\}_{j=0}^{J_1}$ and $CR\left\{t_j^L\right\}_{j=0}^{J_2}$ be their limits in probability as n goes to infinity. Misdiagnosis-related harm can measured by the difference and ratio of the short-term and long-term crude rates. For these two harm measures, we present routine asymptotic normality and confidence interval results as follows.

Asymptotic normality: By central limit theorem and delta method, we have that

$$n^{1/2}\{(\widehat{CR}\{t_j^S\}_{j=0}^{J_1} - \widehat{CR}\{t_j^L\}_{j=0}^{J_2}) - (CR\{t_j^S\}_{j=0}^{J_1} - CR\{t_j^L\}_{j=0}^{J_2})\} \xrightarrow{D} N\left(0, \sigma_{RD}^2\right),$$

$$n^{1/2}\{\log(\widehat{CR}\{t_j^S\}_{j=0}^{J_1} / \widehat{CR}\{t_j^L\}_{j=0}^{J_2}) - \log(CR\{t_j^S\}_{j=0}^{J_1} / CR\{t_j^L\}_{j=0}^{J_2})\} \xrightarrow{D} N\left(0, \sigma_{RR}^2\right),$$

where

$$\sigma_{RD}^2 = \sigma_{\left\{t_j^S\right\}_{j=0}^{J_1}}^2 + \sigma_{\left\{t_j^L\right\}_{j=0}^{J_2}}^2,$$

$$\sigma_{RR}^2 = \frac{\sigma_{\left\{t_j^S\right\}_{j=0}^{J_1}}^2}{CR\left\{t_j^S\right\}_{j=0}^{J_1}} + \frac{\sigma_{\left\{t_j^L\right\}_{j=0}^{J_2}}^2}{CR\left\{t_j^L\right\}_{j=0}^{J_2}},$$

and $\sigma^2_{\{t_j\}_{j=0}^J}$ for a general bin partition using points $\{t_j\}_{j=0}^J$ is the variance of crude rate estimator $\widehat{CR}_{\{t_j\}_{j=0}^J}$. By law of large numbers, delta method, and using the plug-in estimator in places of probabilities, this variance can be estimated by

$$\hat{\sigma}^2_{\{t_j\}_{j=0}^J} = (t_J - t_0)^{-2} \cdot n \cdot \sum_{j=1}^{J} \frac{\sum_{i=1}^n \mathbb{1}(t_{j-1} \le Y_i < t_j,\ \Delta_i\ = 1) \times \sum_{i=1}^n \left\{ \mathbb{1}(Y_i \ge t_{j-1}) - \mathbb{1}(t_{j-1} \le Y_i < t_j,\ \Delta_i\ = 1) \right\}}{\left\{ \sum_{i=1}^n \mathbb{1}(Y_i \ge t_{j-1}) \right\}^3}.$$

Therefore, variances of crude rate difference and ratios can be estimated respectively using

$$\hat{\sigma}^2_{RD} = \hat{\sigma}^2_{\{t_j^S\}_{j=0}^{J_1}} + \hat{\sigma}^2_{\{t_j^L\}_{j=0}^{J_2}},$$

$$\hat{\sigma}^2_{RR} = \frac{\hat{\sigma}^2_{\{t_j^S\}_{j=0}^{J_1}}}{\widehat{CR}_{\{t_j^S\}_{j=0}^{J_1}}} + \frac{\hat{\sigma}^2_{\{t_j^L\}_{j=0}^{J_2}}}{\widehat{CR}_{\{t_j^L\}_{j=0}^{J_2}}}.$$

Note that the variances of crude rate difference and log ratio have clean forms because the crude rates over non-overlapping intervals are in fact independent.

Statistical inference results presented above follow from routine maximum likelihood estimator theories, and construction of the confidence intervals has specifically adopted the delta method. Alternatively, the Fieller's method could be used for the ratio's confidence interval. However, both theoretical[18] and simulation[19] results in the literature have suggested that the Fieller and Delta confidence intervals have only minimal differences, especially when the two quantities in a ratio have a small correlation as is the case in the crude rate ratio.

Confidence intervals: Let $\widehat{RD}$ denote the crude rate difference, $\widehat{RR}$ the crude rate ratio, and let RD and RR be their convergents. The $(1 - \alpha)\%$ confidence intervals can be constructed as

$$(\widehat{RD} - n^{-1/2} \cdot z_{1-\alpha/2} \cdot \hat{\sigma}_{RD}, \widehat{RD} + n^{-1/2} \cdot z_{1-\alpha/2} \cdot \hat{\sigma}_{RD})$$
$$\left(\exp\left\{\log \widehat{RR} - n^{-1/2} \cdot z_{1-\alpha/2} \cdot \hat{\sigma}_{RR}\right\}, \exp\left\{\log \widehat{RR} + n^{-1/2} \cdot z_{1-\alpha/2} \cdot \hat{\sigma}_{RR}\right\}\right),$$

where $z_{1-\alpha/2}$ is the $100 \cdot (1 - \alpha/2)\%$ quantile of standard normal.

Point estimates and confidence intervals can be used to measure the magnitude of harm. However, researchers sometimes also need to qualitatively investigate whether harm exists and whether institutions vary in their diagnostic performances, for which we present a summary of relevant existing hypothesis testing tools that are appropriate under various scenarios.

Testing harm existence: To qualitatively check whether misdiagnosis-related harm exists based on crude rate difference, we test the null hypothesis of RD=0 against the alternative $\widehat{RD} > 0$ using test statistic $\hat{T}_{RD} = n^{1/2} \cdot \widehat{RD}/\hat{\sigma}_{RD}$. Similarly when using crude rate ratio to identify existence of harm, we test the null hypothesis RR=1 against the alternative RR>1 using test statistic $\hat{T}_{RR} = n^{1/2} \log RR/\hat{\sigma}_{RR}$. Test statistics follow standard normal distributions, and calculations of P-value, power, and sample size then follow routine procedures.

Comparing two institutions: We may also be interested in comparing misdiagnosis-related harms across institutions, in which case we would test whether crude rate differences or ratios from two institutions are equal. Specifically, suppose that the crude rate differences are $\widehat{RD}_1$ and $\widehat{RD}_2$ whose corresponding convergents are RD_1 and RD_2. Also denote by $\hat{\sigma}^2_{RD,1}/n_1$ and $\hat{\sigma}^2_{RD,2}/n_2$ the variance estimates for $\widehat{RD}_1$ and $\widehat{RD}_2$ respectively, where n_1 and n_2 are number of patients in the two institutions. We then test the null hypothesis $RD_1 = RD_2$ against the alternative $RD_1 \neq RD_2$ using test statistic $\hat{T}_{RD,1,2} = (\widehat{RD}_1 - \widehat{RD}_2)/\left(\hat{\sigma}^2_{RD,1}/n_1 + \hat{\sigma}^2_{RD,2}/n_2\right)^{1/2}$. Similarly, we can use crude rate ratios to compare institutions. Consider log crude rate ratios $\log \widehat{RR}_1$ and $\log \widehat{RR}_2$ from two institutions. Denote by $\log RR_1$ and $\log RR_2$ their convergents and by $\hat{\sigma}_{RR,1}/n_1$ and $\hat{\sigma}_{RR,2}/n_2$ their variance estimates. We then test the null hypothesis $RR_1 = RR_2$ against the alternative $RR_1 \neq RR_2$ using test statistic $\hat{T}_{RR,1,2} = (\log \widehat{RR}_1 - \log \widehat{RR}_2)/\left(\hat{\sigma}_{RR,1}/n_1 + \hat{\sigma}_{RR,2}/n_2\right)^{1/2}$. Test statistics follow standard normal distributions, and calculations of P-value, power, and sample size then follow routine procedures.

Testing interinstitutional heterogeneity: When researchers compare the diagnostic performances among several institutions, they might not know a priori which two specific institutions to compare. Performing multiple pairwise hypothesis testing creates multiple testing issues, and it could be preferable to test the more general hypothesis of whether harm differs across multiple institutions. For this purpose, we recommend using the following testing procedure of the Wald type.[20] Suppose we have independent crude rate differences $\widehat{RD}_1, \ldots, \widehat{RD}_Q$ from Q institutions indexed by q=1, …, Q, whose limits are RD_1, …, RD_Q. And we assume that variance estimates of these crude rate differences are $\hat{\sigma}^2_{RD,1}/n_1, \ldots, \hat{\sigma}^2_{RD,Q}/n_Q$ respectively, where n_1, …, n_Q are patient sample sizes. Denote by $\widehat{RD}_{\{1,\ldots,Q\}} = (\widehat{RD}_1, \ldots, \widehat{RD}_Q)^\top$, and by $\hat{\Sigma}_{RD,\{1,\ldots,Q\}}$ the variance-covariance matrix estimate of $(\widehat{RD}_1, \ldots, \widehat{RD}_Q)^\top$ such that its diagonal elements are $\hat{\sigma}^2_{RD,q}/n_q$ for q=1, …, Q and its off-diagonal elements are zeros. Formally we are interested in testing the null hypothesis $H_0 : RD_1 = \cdots = RD_Q$ against the alternative that at least one equality does not hold. Let $f(\widehat{RD}_{\{1,\ldots,Q\}}) = (\widehat{RD}_1 - \widehat{RD}_2, \ldots, \widehat{RD}_{Q-1} - \widehat{RD}_Q)^\top$ and we consider test statistic

$$\hat{T}_{RD,\{1,\ldots,Q\}} = f(\widehat{RD}\{1,\ldots,Q\})^{\top} \cdot \left\{ \frac{\partial f(\widehat{RD}\{1,\ldots,Q\})}{\partial \widehat{RD}\{1,\ldots,Q\}} \hat{\Sigma}_{RD,\{1,\ldots,Q\}} \left(\frac{\partial f(\widehat{RD}\{1,\ldots,Q\})}{\partial \widehat{RD}\{1,\ldots,Q\}} \right)^{\top} \right\}^{-1} \cdot f(\widehat{RD}\{1,\ldots,Q\}),$$

which follows chi-squared distribution with degree of freedom $Q-1$. We can similarly test whether levels of harm differ across Q institutions in terms of log crude rate ratios, denoted as $\log \widehat{RR}_1, \ldots, \log \widehat{RR}_Q$. Let $\hat{\Sigma}_{RR,\{1,\ldots,Q\}}$ be the variance-covariance matrix estimate for $\log \widehat{RR}_{\{1,\ldots,Q\}} = \left(\log \widehat{RR}_1, \ldots, \log \widehat{RR}_Q\right)^{\top}$. We test the null hypothesis $H_0 : RR_1 = \cdots = RR_Q$ against the alternative that at least one equality does not hold. We consider the test statistic

$$\hat{T}_{RR,\{1,\ldots,Q\}} = f(\log \widehat{RR}\{1,\ldots,Q\})^{\top} \cdot \left\{ \frac{\partial f(\log \widehat{RR}\{1,\ldots,Q\})}{\partial \log \widehat{RR}\{1,\ldots,Q\}} \hat{\Sigma}_{RR,\{1,\ldots,Q\}} \left(\frac{\partial f(\log \widehat{RR}\{1,\ldots,Q\})}{\partial \log \widehat{RR}\{1,\ldots,Q\}} \right)^{\top} \right\}^{-1} \cdot f(\log \widehat{RR}\{1,\ldots,Q\}),$$

which also follows chi-squared distribution with degree of freedom $Q-1$. Calculations of P-value, power, and sample size then follow routine procedures, and the order of institutions does not affect any result.

4 | SIMULATION STUDIES

In this section, we present simulation studies that investigate bias and performance of the crude-rate-based harm estimators in detecting misdiagnosis-related harm compared with those of Nelson-Aalen estimator. Through these simulations, we obtain an understanding of the scenarios, in terms of number of events, level of censoring, and sample size, under which multibin crude rates with small number of bins, especially the single-bin crude rate, have little bias in measuring harm and are equally powerful in detecting harm.

We generate time of event incidences following Weibull distributions (scaled by a constant of three). We take shape parameter of the Weibull distributions from $k = 1, 0.8, 0.6$, representing scenarios with no harm, mild harm, and severe harm. Size parameters of the Weibull distributions are chosen such that the number of expected incidences over short-term window (0,1.5] is 5, 30, and 200 respectively for sample sizes 1000, 10 000, and 100 000. As a result, number of expected incidences over the long-term window [3,6) may vary. We also consider different levels of uniform censoring at 10%, 35%, and 60% over the union of the short-term window (0,1.5] and the long-term window [3,6). That is, the probability of observing a censored incidence in (0,1.5]∪[3,6) is 10%, 35%, and 60% respectively. These simulation scenarios emulate real data cases with different event rates, censoring rates, and sample sizes. A variety of scenarios exist outside of what are considered in our simulations, for example, censoring distributions can be nonuniformâ–we do not attempt to be comprehensive but hope to provide some typical examples that practitioners can use as a guide, a case study that they can extrapolate from, or a reference for how to perform similar simulations in a different context. See Table 1 for the specification of simulation parameters as well as the true values of cumulative hazard differences and ratios under various scenarios.

We generate data with 1000 replications for each scenario and study the crude rate differences and ratios calculated for intervals [0,1.5) and [3,6) based on multibin crude rate estimators with 1, 2, 4, 8, 16, or 32 equal-sized bins per 1.5 unit of time. As a reference, we also calculate the Nelson-Aalen estimates. Empirical biases compared to the difference or log ratio of average cumulative hazards, as well as type-I error and power for detecting the existence of misdiagnosis-related harm, are summarized in Figures 1 and 2 for censoring at 10% and in additional figures in the Supplementary Information for censoring at 35% and 60%.

We observe that the level of bias decreases with an increased number of bins and flattens with around 10 bins, indicating that further increasing the number of bins used for J-bin crude rate calculation might not add much improvement. For hypothesis testing, we observe close to theoretical type-I error throughout simulated scenarios and observe that the power for detecting misdiagnosis increases with an increase in the expected number of incidences in the short-term window. On the other hand, power appears to be almost invariant to the number of bins used in simulated scenarios. We do observe some spurious results when the event rate is high under heavy censoring at 60%. But in general, these results suggest that using a small number of bins is often sufficient for appropriately detecting misdiagnosis-related harm despite the potential bias where the event rate and censoring rate are low as investigated in our simulations. As expected, we observe smaller bias and higher hypothesis testing power with increasing sample size.

Note that we have taken equal-sized bins without considering how many incidences or censoring each bin aggregates. Theoretically, what directly affects the accuracy of estimation and efficiency of testing is the number of events and censoring observed in each bin instead of the number of bins. Therefore, a theoretically more efficient approach is to aggregate equal numbers of incidences or censoring across bins. However, adopting such time bins contradicts the point of using a simple measure, and therefore using equal-sized bins is more practical. As a rule of thumb, 10 bins for crude rate calculations are often sufficient for minimizing the bias and maintaining the power for harm detection. We can also utilize simulation results to provide a more specific guide on how many bins to use by finding the simulation scenarios that are close to real data in terms of number of events, censoring rate, and sample size. We illustrate this strategy in Section 5.

5 | DATA ANALYSIS

In this section, we illustrate the proposed analysis procedure with analyses on the Longitudinal Health Insurance Database (LHID). The LHID is an one-million patient database randomly sampled from the National Health Insurance Research Database of Taiwan (NHIRD). The Taiwan NHIRD is an electronic medical record database that dates back to March 1995 and covers 99% of Taiwan’s populations with close-to-none missing data or out-of-network patients. In this analysis, we include 144 355 patients who were enrolled in the Taiwan National Health Insurance (NHI) program in 2010. We then retrospectively identify patients who had a dizziness visit in outpatient departments between 1 January 2002 and 31 December 2009. We excluded patients that were admitted or referred to emergency departments at the dizziness visits because they were considered to have received "positive" diagnoses while we only aim to measure the harm of false negative diagnoses. The included 144 355 patients remained under observation for 365 days for the first stroke inpatient hospitalization after their first dizziness visit within the time frame under consideration. By including only those patients who were enrolled in 2010, we implicitly exclude patients who had died before 2010, creating a potentially biased sample from the general population. However, addressing this issue is beyond the scope of this article, and we would restrict any conclusion of the data analyses to the LHID sampled population.

We analyze stroke event time during the short-term window from 1 to 30 days and during the long-term window from 91 to 180 days after the index visit. The short-term window of 1 to 30 days is chosen because it has been considered in previous literature as a window for defining potentially missed stroke diagnosis.[10] Also the window is wide enough for sufficient number of observed events. The long-term window from 91 to 180 days is chosen such that the stroke risk has mostly flattened and that the window is wide enough for sufficient observations.

We observe 340 stroke events within 30 days after the index visit, and 168 events from 91 to 180 days. The data most closely resembles simulation scenarios with expected number of short-term events around 200, that of long-term events around 100, sample size 100 000, and censoring at 10%. Simulation results for such scenarios indicate that using the single-bin crude rate is sufficient for small bias, an appropriate type-I error, and a high hypothesis testing power. Thus we calculate the single-bin crude rate difference and ratio to quantify the misdiagnosis-related harm.

We estimate the one-bin crude rate difference to be 18.94 cases per 30 days and per 10 000 people (95% confidence interval [16.45,21.43], P-value < .01), suggesting the existence of misdiagnosis-related harm. We also estimate the one-bin crude rate log ratio to be 1.78 (95% confidence interval [1.59,1.96], P-value < .01), which supports the same conclusion.

To illustrate the proposed hypothesis testing tools, we study age disparity in misdiagnosis-related harm by categorizing patients into young (age≤40, 49 872 patients), middle-aged (40<age≤60, 54 491 patients), and old (age>60, 39 992 patients) groups. We then test the existence of group heterogeneity using hypothesis testing procedures discussed in Section 3.2. Results suggest that misdiagnosis-related harm is different among age groups on the absolute scale (in terms of crude rate difference; chi-squared statistic 3×10^{10}, P-value < .01), but not on the relative scale (in terms of crude rate ratio; chi-squared statistic 0.31, P-value .96). These results suggest that the different levels of harm incurred in different age groups are likely due to lower stroke incidences in younger groups rather than diagnostic disparities. If researchers are interested in comparing two specific groups, further hypothesis testing can be carried out.

To further check the validity of using the single-bin crude rates instead of the more complicated multibin versions, we perform harm quantification analyses using K-bin crude rates for $K=1, \ldots, 15$. We present the crude rate difference and ratio estimation and hypothesis testing results concerning age heterogeneity in Figure 3. We observe that crude rate differences and log ratios have almost identical estimates and confidence intervals across different numbers of bins, and hypothesis testing results are barely affected. These results confirm that choosing to use one bin based on simulation results has been a valid strategy in terms of obtaining equally informative harm measures.

6 | DISCUSSION

In this article, we provided thorough discussions on statistical tools and insights related to the easy-to-compute crude rate and crude-rate based measures for timely monitoring misdiagnosis-related harm. We generalized the single-bin crude rate to its multibin version and provided an interpretation for using crude rate difference and ratio as harm quantification, linking the crude rate to cumulative hazard. We then analyzed the deviation of crude rate from the Nelson-Aalen estimator which statisticians would have used for estimating the cumulative hazard if the more sophisticated analyses were feasible. The deviation was shown to be small in the presence of low event and censoring rates. We further illustrated with simulation studies and real data analysis using stroke event occurrences from the LHID that a small number of bins, or even a single bin, was often sufficient for an estimate with minimal bias and powerful hypothesis testing results.

Specially, we observed in our simulation studies that the single-bin crude rate was almost as accurate and efficient as its multibin or Nelson-Aalen-estimator based alternatives when short-term and long-term event rates were below 10% per unit of time and when the censoring rate was below 10% per unit of time. This translates to no more than 100 events and 100 censoring per 1000 patients between 1 and 30 days and no more than 300 events and 300 censoring per 1000 patients between 91 and 180 days, which are usually satisfied in the misdiagnosis context with administrative data.

We also observed that for all simulated scenarios, using a 10-bin crude rate would yield results close to those based on the Nelson-Aalen estimator. For other scenarios that researchers may encounter in practice, our simulation studies can be used as a guide, a case study to extrapolate from, or a reference on how such simulations can be run.

Our work faces a few limitations. First, the interpretation of crude rate difference and ratio for quantifying misdiagnosis-related harm was established with approximation under the assumption of a low event rate, which is common in the context of diagnostic errors but might not be as common otherwise. Second, partitions with irregular time points were not studies although they might reduce bias. Third, the choice of the short-term and long-term windows was not discussed despite its relevance—using a wide short-term window improves robustness but might dilute signal, while using a long-term window that is wide and far from the short-term window is desirable but not always practical as longer follow-up is needed. There are trade-offs that take thoughtful consideration in choosing these windows. However, overcoming these limitations likely comes at a price of complexity and is beyond the scope of this article. On the other hand, more data-driven ways of selecting the short-term and long-term windows is one exciting direction for future work.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

This work was supported by a grant from the Gordon & Betty Moore Foundation (5756) and the Armstrong Institute Center for Diagnostic Excellence. Dr Wang's effort is partially supported by NIH grant R01AG068002. Dr Liberman's effort is supported by National Institutes of Health grant K23NS107643.

Funding information

Gordon & Betty Moore Foundation, Grant/Award Number: 5756; National Institutes of Health, Grant/Award Number: R01AG068002; National Institutes of Health, Grant/Award Number: K23NS107643

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

REFERENCES

1. National Academies of Sciences, Engineering, and Medicine. Improving Diagnosis in Health Care. Washington, DC: National Academies Press; 2015.
2. Singh H, Giardina TD, Meyer AND, Forjuoh SN, Reis MD, Thomas EJ. Types and origins of diagnostic errors in primary care settings. JAMA Internal Med. 2013;173(6):418–425. [PubMed: 23440149]
3. Singh H, Meyer AND, Thomas EJ. The frequency of diagnostic errors in outpatient care: estimations from three large observational studies involving US adult populations. BMJ Qual Saf. 2014;23(9):727–731.
4. Hayward RA, Hofer TP. Estimating hospital deaths due to medical errors: preventability is in the eye of the reviewer. JAMA. 2001;286(4):415–420. [PubMed: 11466119]
5. Luck J, Peabody JW, Dresselhaus TR, Lee M, Glassman P. How well does chart abstraction measure quality? a prospective comparison of standardized patients with the medical record. Am J Med. 2000;108(8):642–649. [PubMed: 10856412]
6. Kerber KA, Morgenstern LB, Meurer WJ, et al.Nystagmus assessments documented by emergency physicians in acute dizziness presentations: a target for decision support?Acad Emerg Med. 2011;18(6):619–626. [PubMed: 21676060]
7. Kim AS, Fullerton HJ, Johnston S. Claiborne. risk of vascular events in emergency department patients discharged home with diagnosis of dizziness or vertigo. Ann Emerg Med. 2011;57(1):34–41. [PubMed: 20855127]
8. Royl G, Ploner CJ, Leithner C. Dizziness in the emergency room: diagnoses and misdiagnoses. Eur Neurol. 2011;66(5):256–263. [PubMed: 21986277]
9. Lee C-C, Ho H-C, Su Y-C, et al.Increased risk of vascular events in emergency room patients discharged home with diagnosis of dizziness or vertigo: a 3-year follow-up study. PloS One. 2012;7(4):e35923. [PubMed: 22558272]
10. Newman-Toker DE, Moy E, Valente E, Coffey R, Hines AL. Missed diagnosis of stroke in the emergency department: a cross-sectional analysis of a large population-based sample. Diagnosis. 2014;1(2):155–166. [PubMed: 28344918]

11. Atzema CL, Grewal K, Lu H, Kapral MK, Kulkarni G, Austin PC. Outcomes among patients discharged from the emergency department with a diagnosis of peripheral vertigo. Ann Neurol. 2016;79(1):32–41. [PubMed: 26385410]

12. Madsen TE, Khoury J, Cadena R, et al.Potentially missed diagnosis of ischemic stroke in the emergency department in the Greater Cincinnati/Northern Kentucky Stroke Study. Acad Emerg Med. 2016;23(10):1128–1135. [PubMed: 27313141]

13. Liberman AL, Newman-Toker DE. Symptom-Disease Pair Analysis of Diagnostic Error (SPADE): a conceptual framework and methodological approach for unearthing misdiagnosis-related harms using big data. BMJ Qual Saf. 2018;27(7):557–566.

14. Mane KK, Rubenstein KB, Nassery N, et al.Diagnostic performance dashboards: tracking diagnostic errors using big data. BMJ Qual Saf. 2018;27(7):567–570.

15. Nelson WTheory and applications of hazard plotting for censored failure data. Technometrics. 1972;14(4):945–966.

16. Aalen ONonparametric inference for a family of counting processes. Ann Stat. 1978;6(4):701–726.

17. Winnett A, Sasieni P. Adjusted nelson–aalen estimates with retrospective matching. J Amer Stat Assoc. 2002;97(457):245–256.

18. Hole AR. A comparison of approaches to estimating confidence intervals for willingness to pay measures. Health Econom. 2007;16(8):827–840.

19. Beyene J, Moineddin R. Methods for confidence interval estimation of a ratio parameter with application to location quotients. BMC Med Res Methodol. 2005;5(1):1–7. [PubMed: 15636638]

20. Wald ATests of statistical hypotheses concerning several parameters when the number of observations is large. Trans Am Math Soc. 1943;54(3):426–482.

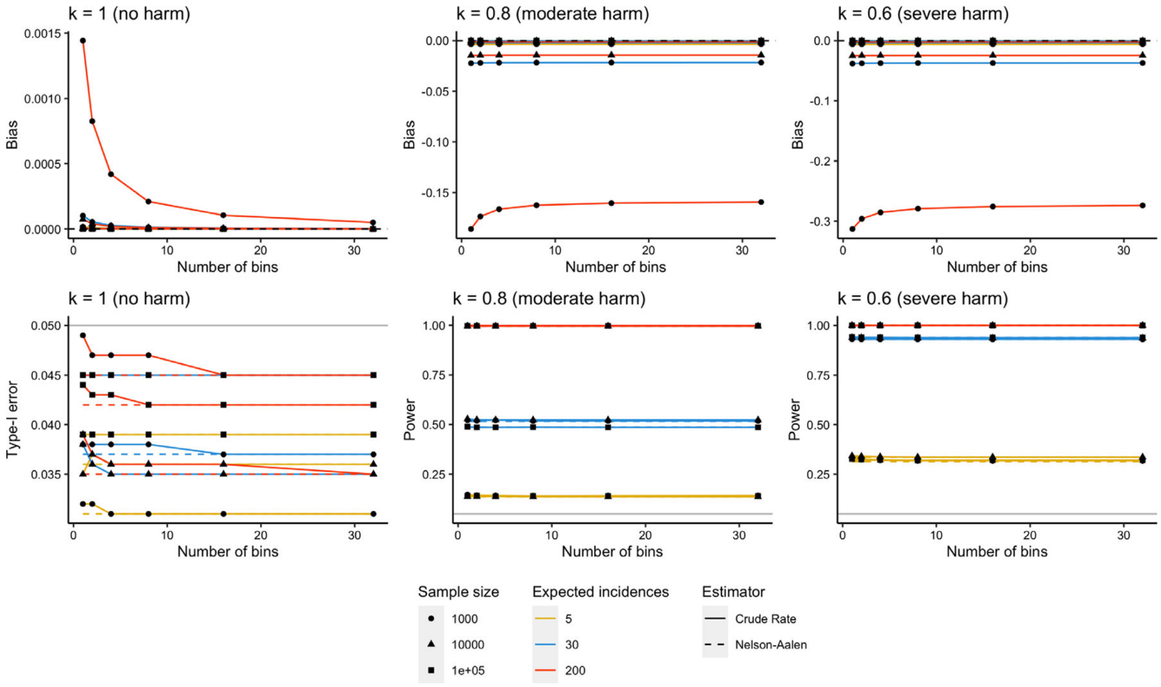

FIGURE 1.

Simulation results for crude rate difference estimation and hypothesis testing when we have 10% censoring. Biases decrease with increased number of bins, decreased number of expected events, and larger sample sizes; type-I errors are generally all close to 0.05; power increases with more expected incidences, but not much with increase in sample size or number of bins [Colour figure can be viewed at wileyonlinelibrary.com]

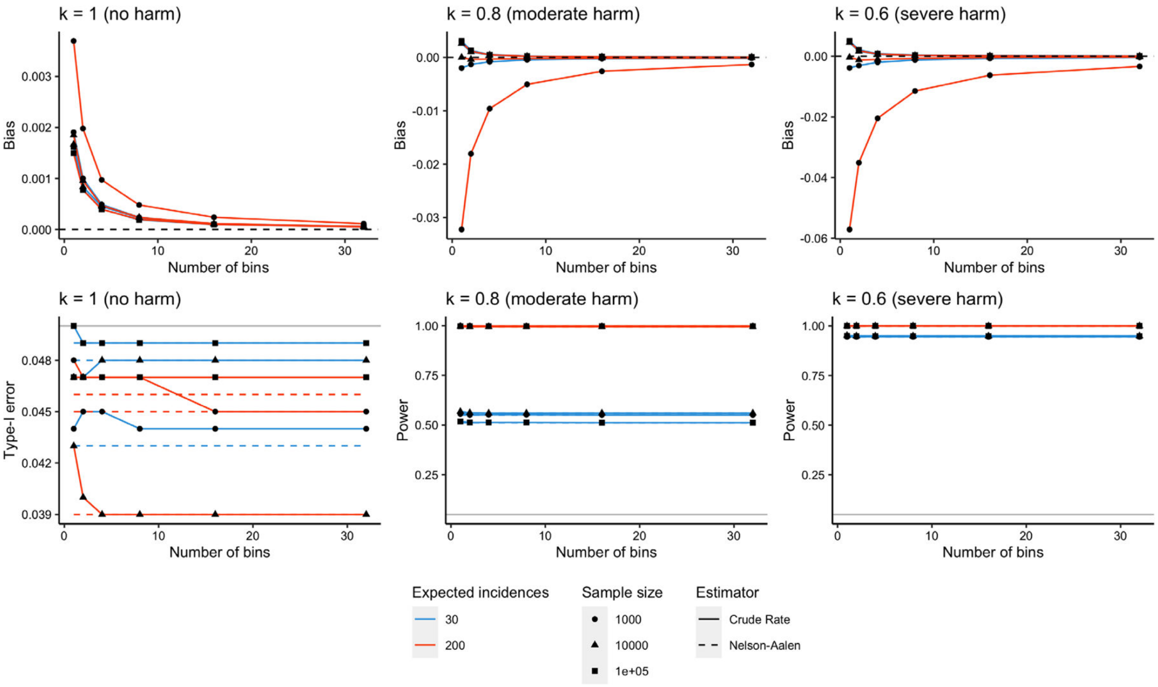

FIGURE 2.

Simulation results for log crude rate ratio estimation and hypothesis testing when we have 10% censoring. Biases decrease with increased number of bins, decreased number of expected events, and larger sample sizes; type-I errors are generally all close to 0.05; power increases with increased numbers of expected incidences, but not much with increase in sample size or number of bins [Colour figure can be viewed at wileyonlinelibrary.com]

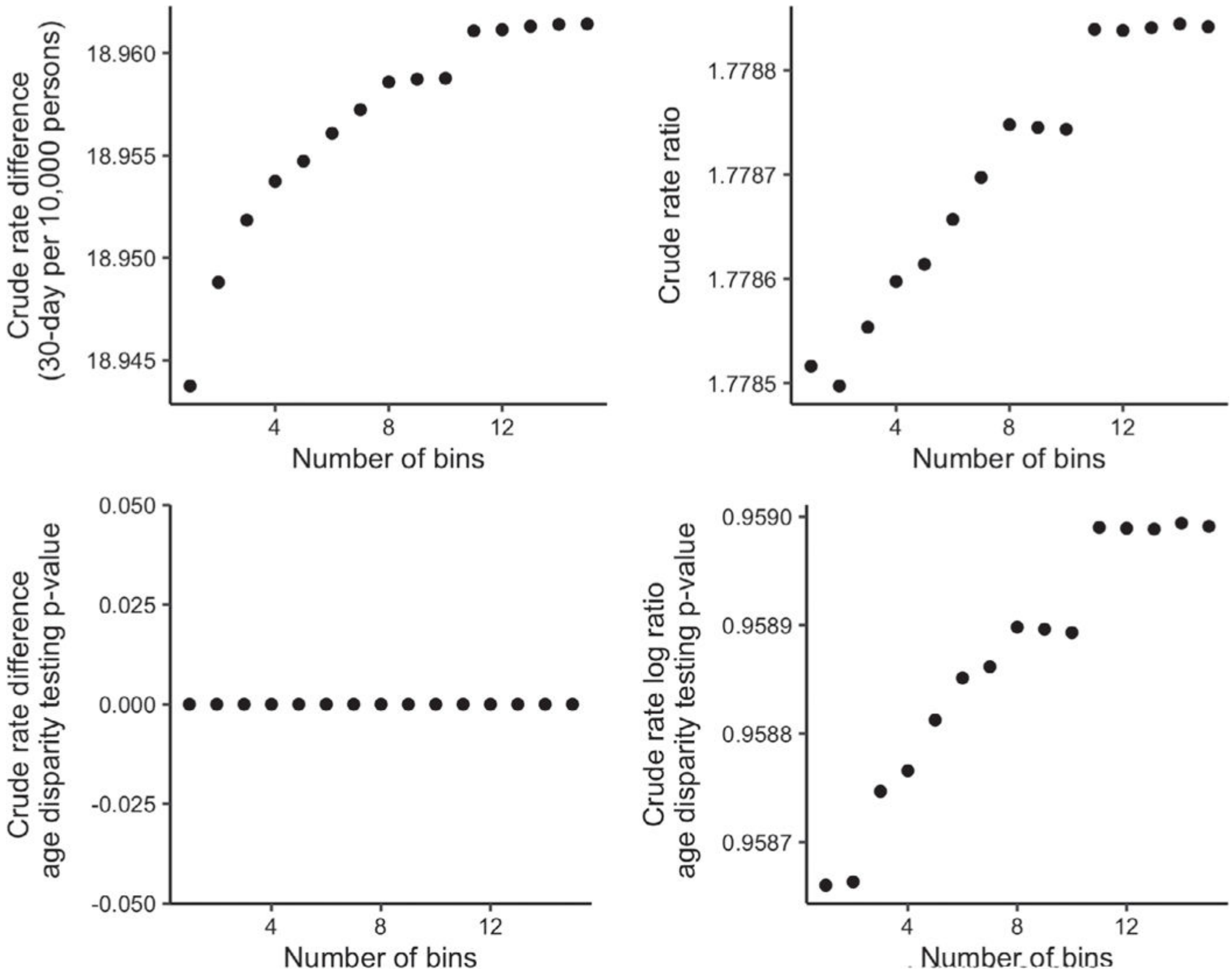

FIGURE 3.

Comparison of analysis results of the Taiwan NHIRD using difference number of bins. We observe that increasing the number of bins from 1 to 15 results in less than 0.2% of increase in crude rate difference, less than 0.03% of increase in crude rate ratio, and ignorable changes in test *P*-values. These observations suggest that using a small number of bins has little effect on analysis results

TABLE 1

Simulation settings

k	**$\log_{10}n$**	***λ***	**Expected number of incidences in (0,1.5]**	**Expected number of incidences in [3,6)×0.5**	**Average cumulative hazard difference×*n***	**Average cumulative hazard log ratio**
0.6	3	3405.7	5	1.9	6.11	0.939
		168.31	30	11.2	37.1	0.939
		6.09	200	57.1	272	0.939
	4	158 673.96	5	2	6.09	0.939
		7992.47	30	11.7	36.6	0.939
		333.65	200	76	246	0.939
	5	7 367 755.99	5	2	6.09	0.939
		371 814.22	30	11.7	36.6	0.939
		15 722.76	200	77.9	244	0.939
0.8	3	374.88	5	3.2	3.56	0.438
		39.29	30	18.3	21.6	0.438
		3.26	200	84.8	158	0.438
	4	6685.31	5	3.2	3.55	0.438
		710.81	30	19.2	21.3	0.438
		65.65	200	124.2	143	0.438
	5	118 917	5	3.2	3.55	0.438
		12 661.55	30	19.3	21.3	0.438
		1180.7	200	128.5	142	0.438
1	3	99.75	5	4.9	0	0
		16.42	30	27.8	0	0
		2.24	200	115.2	0	0
	4	999.75	5	5	0	0
		166.42	30	29.8	0	0
		24.75	200	190.2	0	0
	5	9999.75	5	5	0	0
		1666.42	30	30	0	0
		249.75	200	199	0	0