CBER SENTINEL METHODS

QUANTITATIVE BIAS ANALYSIS METHODOLOGY DEVELOPMENT: SEQUENTIAL BIAS ADJUSTMENT FOR OUTCOME MISCLASSIFICATION

FINAL REPORT

Prepared by: Chandrasekar Gopalakrishnan, MD, MPH,[1] Timothy L Lash, DSc, MPH,[2] Richard A Forshee, PhD,[3] Noelle Cocoros, DSc, MPH,[4] Yun Lu, PhD,[3] Sandra Feibelmann, MPH,[4] Christopher Jankosky, MD, MPH,[3] Martin Kulldorff, PhD,[1] David McClure, PhD,[5] Matthew P Fox, DSc, MPH,[6] Joshua J Gagne, PharmD, ScD[1]

Author Affiliations: 1. Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA; 2. Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, GA; 3. Office of Biostatistics and Epidemiology, Center for Biologics Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD; 4. Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, MA; 5. Center for Clinical Epidemiology & Population Health, Marshfield Clinic Research Institute, Marshfield, WI; 6. Departments of Epidemiology and Global Health, School of Public Health, Boston University, Boston, MA

August 4, 2017

Sentinel is a project sponsored by the U.S. Food and Drug Administration (FDA) to monitor the safety of FDA-regulated medical products. Sentinel is one piece of the Sentinel Initiative, a multi-faceted effort by the FDA to develop a national electronic system that complements previously existing methods of safety surveillance. Sentinel Collaborators include Data and Academic Partners that provide access to health care data and ongoing scientific, technical, methodological, and organizational expertise. The Sentinel Coordinating Center is funded by the FDA through the Department of Health and Human Services (HHS) (Contract number HHSF223201400030I).

CBER Sentinel Methods

Quantitative Bias Analysis Methodology Development: Sequential Bias Adjustment for Outcome Misclassification

Final Report

Table of Contents

I. BACKGROUND AND OBJECTIVE

As with all observational analyses that evaluate associations between medical products and health outcomes, Sentinel assessments are susceptible to potential biases, including those due to confounding and misclassification.[1] In particular, most Sentinel assessments rely on administrative claims data to ascertain exposures and outcomes and claims-based outcome definitions typically lack perfect sensitivity, specificity, or both, which can lead to bias.[2] In particular, administrative claims are transactional records that may not always capture an outcome if, for example, the outcome did not lead to medical attention or may suggest any outcome when one did not actually occur, such as when providers make rule-out diagnoses. Medical record review is often used to confirm outcomes identified in the claims data.[2]

When conducted within the context of a specific medical product-outcome assessment, medical record validation can be used to adjudicate events with the intent of including only confirmed cases in the analysis, similar to event adjudication in clinical trials. Outside of a given assessment, medical record validation can be used to evaluate the performance of a claims-based outcome definition, such as by estimating the positive predictive value (PPV) of the algorithm. Quantitative bias analysis (QBA) methods quantify residual bias in effect estimation by modeling structural assumptions about the mechanisms that can cause bias.[3] To adjust for bias due to outcome misclassification, QBA can use measures of performance of the claims-based definition derived either from a prior validation study or from a sample of the outcomes identified by the claims-based algorithm within the assessment of interest. As medical record validation can be costly and time-consuming, QBA enables investigators to address bias due to outcome misclassification much in the same way as would adjudicating all events, but potentially in much less time and at a lower cost. However, since Sentinel is often interested in expedited evidence generation and since time and monetary resources are finite, further reducing the time and cost associated with medical record validation is important.

The objective of this project was to develop an approach to using QBA to adaptively determine whether and when medical chart validation to inform adjustment for bias due to outcome misclassification can be stopped early because additional chart review would not change the inference from a given Sentinel assessment.

A. METHODS DEVELOPMENT

As important context, results of medical chart validation are typically returned to the Sentinel Operations Center (SOC) in batches over a period of time. Depending on the number of charts to be reviewed, the time between receipt of first and last results can span many months. One potential use of the proposed method is to determine when the return of results can be stopped early. Alternatively, Sentinel's Distributed Data Network, in which individual-level data remain behind a firewall at each Data Partner (DP), creates another partition along which decisions to stop or continue with medical chart validation could be made. For example, the medical chart validation process could be initiated at one DP and, once results are returned from this DP, the method could be applied to determine whether medical chart validation results from a second DP are needed, before initiating the validation process within the second DP.

The approach to methods development for this project involved four stages. In the first stage, the Workgroup conducted a literature review to determine whether similar methodological approaches had been previously proposed and to develop a deeper understanding of current and best practices of medical chart validation methods and processes. The second stage involved the identification and

review of the necessary input parameters for conducting a QBA for addressing bias due to outcome misclassification. Once the optimal validation design and QBA input parameters were identified, the third phase was to create a blueprint for the method using a hypothetical data example. Finally, in the fourth phase, we applied the method to a real data example. The remainder of this report is structured according to these four stages.

B. STAGE 1: LITERATURE REVIEW

The Workgroup initially conducted a review of the literature with three specific aims: (1) to determine whether similar methodological approaches had been previously proposed or used; (2) to better understand current and best practices for medical chart validation methods and processes used in Sentinel assessments and, more broadly, in pharmacoepidemiology at large; and (3) to review methodological papers that evaluated or compared different designs for the conduct of medical chart validation. Below, we briefly summarize the main findings of the literature review.

The literature review did not identify any proposals or uses of similar approaches to the method proposed by the Workgroup. The review also revealed that medical chart validation studies in the literature regularly do not specify clear methodology in requesting charts for review, including the strategy for selecting the number of charts to be reviewed. When samples of charts were obtained for outcome validation, they were typically a random sample of all cases within a study cohort with the sampling fraction usually determined arbitrarily or by logistical and financial constraints and without regard to exposure status when they were conducted within the context of a given exposure-outcome study.[4 5 6 7 8 9 10 11] In a methodological paper published by Holcroft *et al*. comparing the performance of different validation designs using a fixed sampling fraction, balanced sampling designs outperformed random sampling designs.[12] With a fixed number of outcomes to be validated, a balanced design suggests sampling an equal number of exposed and unexposed outcomes in order to maximize precision in estimation of the PPV in both exposure groups. In contrast, random sampling of outcomes without respect to exposure status would result in an imprecise PPV estimate among exposed cases in settings with an uncommon exposure. The Workgroup determined that the proposed methodology should be built on the balanced design.

QBA to address bias due to outcome misclassification requires estimates of either sensitivity and specificity or PPV and negative predictive value (NPV). However, outcome validation studies in Sentinel and in pharmacoepidemiology typically estimate only PPV because they start with the outcomes identified by the claims-based definition and determine what proportion are true cases based on medical charts. Given that many health outcomes of interest in Sentinel assessments and in pharmacoepidemiology studies are uncommon, obtaining reliable estimates of sensitivity, specificity, or NPV would typically require review of a prohibitively large number of medical charts. As discussed below in stage 2, assumptions about NPV are necessary in order to perform QBA when only information on PPV is available.

C. STAGE 2: IDENTIFICATION AND REVIEW OF THE PARAMETERS NECESSARY TO PERFORM QUANTITATIVE BIAS ANALYSES

In this section, we first introduce simple sensitivity analysis in which one fixed value is assigned to each parameter of a bias model to yield a single revised estimate of association.[3 13] We then extend the discussion to probabilistic bias analysis (PBA), which is the QBA technique used in the proposed method. To illustrate simple sensitivity analysis, we will use the following data in **Figure 1** from a hypothetical claims-based study examining the association between some exposure and some outcome. The

hypothetical study included 1,000 exposed individuals and 1,000 unexposed individuals. The claims-based outcome definition identified outcomes in 40 exposed individuals and 20 unexposed individuals. The resulting observed risk ratio (RR) was 2.00 with a 95% confidence interval (CI) from 1.18 to 3.40.

Figure 1. Data from a hypothetical exposure-outcome study

		Exposure +	Exposure -
Outcome	+	40	20
	-	960	980

Imagine that we conducted medical record validation using a balanced design by sampling 10 exposed and 10 unexposed outcomes and let us assume that we confirmed 9 out of 10 exposed outcomes and 8 out of 10 unexposed outcomes as true cases. The corresponding PPVs, therefore, would be 90% among exposed cases and 80% in unexposed cases. By multiplying the number of outcomes identified using the claims-based definition in each exposure group by the respective PPV, we can obtain a bias-adjusted data table (**Figure 2**). In the bias-adjusted table, we can see that 36 of the 40 exposed outcomes were modeled as true cases and 16 of the 20 unexposed cases were modeled as true cases. The bias-adjusted RR is 2.25.

Figure 2. Simple sensitivity analysis to adjust for imperfect PPV

		Exposure E+	Exposure E-
Outcome	+	40	20
	-	960	980
		*(40*0.90)*	*(20*0.80)*
		Exposure	
		E+	**E-**
Outcome	+	36	16
	-	964	984

While this simple sensitivity analysis is intuitive and straightforward, it does not account for random error in the PPV estimates from the validation or random error from the original safety assessment. Rather than using only the PPV point estimate, PBA allows for the incorporation of both sources of random error by sampling from a distribution of PPV and RR estimates. This process of incorporating uncertainty will be described in more detail in subsequent sections.

As mentioned above, the usual approach for medical record validation only yields estimates of PPV. In this hypothetical example, we assumed an NPV of 100%. An NPV will be 100% only when the sensitivity of a claims-based outcome definition is also 100%; that is, when the algorithm identifies every true case. While claims-based definitions are likely not 100% sensitive, NPV will typically be very high in Sentinel

assessments when the health outcome of interest is uncommon, even when sensitivity is fairly low. For example, when the outcome incidence is 1 in 1,000, an algorithm with only 50% sensitivity would have an NPV of 99.95% (assuming specificity of 99%). To evaluate the impact of imperfect NPV, the Workgroup conducted a series of simulations, which are summarized in the Appendix. Overall, the simulation results suggested that, when specificity and outcome prevalence are low, bias can be large even with perfect sensitivity and NPV. The results also suggested that, at a given specificity and outcome prevalence, differences in sensitivity between exposure groups can have a large impact on absolute bias even with very high NPV values. Based on these findings, the Workgroup recommends that, while assumptions about high NPV may be tenable, NPV values should be varied within the proposed method to reflect potential differences in sensitivity between exposure groups.

As an alternative to making assumptions about NPV, one could use the PPV to estimate the specificity of the claims-based outcome definition. However, this approach would still require an assumption about sensitivity and would require an additional assumption about the true incidence of the outcome within each exposure group in order to convert from PPV to specificity. If the true outcome incidence were known within each exposure group, the exposure-outcome assessment would not be necessary. Thus, the proposed method utilizes PPV for the outcome definition across exposure groups, as provided by the balanced design, and makes assumptions about NPV.

D. STAGE 3: DEVELOPING A FRAMEWORK TO CONDUCT ADAPTIVE CHART REVIEW AND APPLICATION TO A HYPOTHETICAL EXAMPLE

Building on the findings from the first two phases described above, the Workgroup developed a 7-step framework for the sequential application of PBA methods to the medical chart validation process to adjust for outcome misclassification within a Sentinel assessment, with the intent of determining when additional chart validation results are no longer needed (i.e. when medical chart review can be stopped). The steps in the framework are outlined below in the context of a given Sentinel assessment and illustrated in **Figure 3**.

Step 1 – Planning phase: Step 1 involves initial planning related to the method before the conduct of a given safety assessment, with decisions about the objective of the validation and thresholds on when to stop chart review. For example, this step may include the following: identify prior outcome validation studies to obtain initial PPV estimates, begin developing assumptions for other parameters (e.g., NPV), discuss and agree upon objective of validation and threshold for continuing versus stopping medical chart validation, identify sensitivity analyses and relevant parameters and how results of sensitivity analyses will be used in the decision to continue or stop. For example, the objective of the validation might be to rule out a bias-adjusted estimate with a 95% simulation interval that includes 1.00. The stopping threshold is the probability of meeting the objective required in order to have sufficient confidence to terminate the validation process.

Step 2 – Conduct safety assessment: In Step 2, the safety assessment is conducted in Sentinel using all outcomes identified in the electronic data from the claims-based outcome definition. An initial effect estimate is obtained.

Step 3 – Assess utility of and decide whether to initiate validation study: At Step 3, an initial PBA is performed to estimate the range of PPVs in each exposure group that would be needed to dismiss outcome misclassification as an explanation for the findings. A decision is then made about whether the medical chart validation will provide value. For example, given certain results from the safety assessment and beliefs about the PPV, perhaps based on the

findings of Step 1, it may be determined that outcome misclassification cannot possibly or is highly unlikely to explain the safety assessment results. In such cases, medical chart validation would not be started.

Step 4 – Initiate validation study: If the decision is made in Step 3 to proceed with medical chart validation, charts of all outcomes identified by the claims-based algorithm, or for a sample thereof based on the balanced design, would be requested from one or more DPs. As is standard practice in Sentinel, the DP(s), or a contracted vendor, identifies the individual patients and sends requests to the respective healthcare providers for the charts. The DP(s) or a contracted vendor redacts the charts of protected health information (PHI) except for dates. The redacted charts are typically sent to the SOC to confirm that the correct charts were received and to ensure PHI was redacted. The redacted charts are then sent to an adjudicator who reviews them and sends the validation results to the SOC in batches.

Step 5 – Sequential adjustment of effect estimate from initial safety assessment: There are three main activities performed as part of Step 5:

- **Batching of validation results**: As described above, batching can occur either over a period of time, by DP, or both. With each batch, the cumulative PPV among exposed outcomes (PPV E+) and unexposed outcomes (PPV E-) are estimated.

- **Determine input bias parameters for PBA**: With each batch, the estimated cumulative PPVs, and measures of their uncertainty, are used as input parameters for PBA. NPVs or a range of NPVs for each exposure group are assumed and entered as inputs into the PBA.

- **Perform sequential PBA adjustment**: With each batch, PBAs are performed by randomly sampling PPVs based on the cumulative PPVs to sequentially adjust for outcome misclassification as the results from the medical chart validation become available. Bias-adjusted estimates are obtained by taking the median point estimate from the PBA and a 95% simulation interval is obtained by taking the estimates at the 2.5^{th} and 97.5^{th} percentiles of the PBA point estimate distribution. The simulation interval incorporates both the uncertainty in the bias parameters and of the initial safety assessment. For this project, publicly available software was used to conduct the PBAs.[3]

Step 6 – Determine whether the pre-defined threshold is met: With each sequential adjustment at each batch, the probability *(ζ)* of observing a bias-adjusted RR with a simulation interval that meets a given criterion of interest is estimated by the following steps.

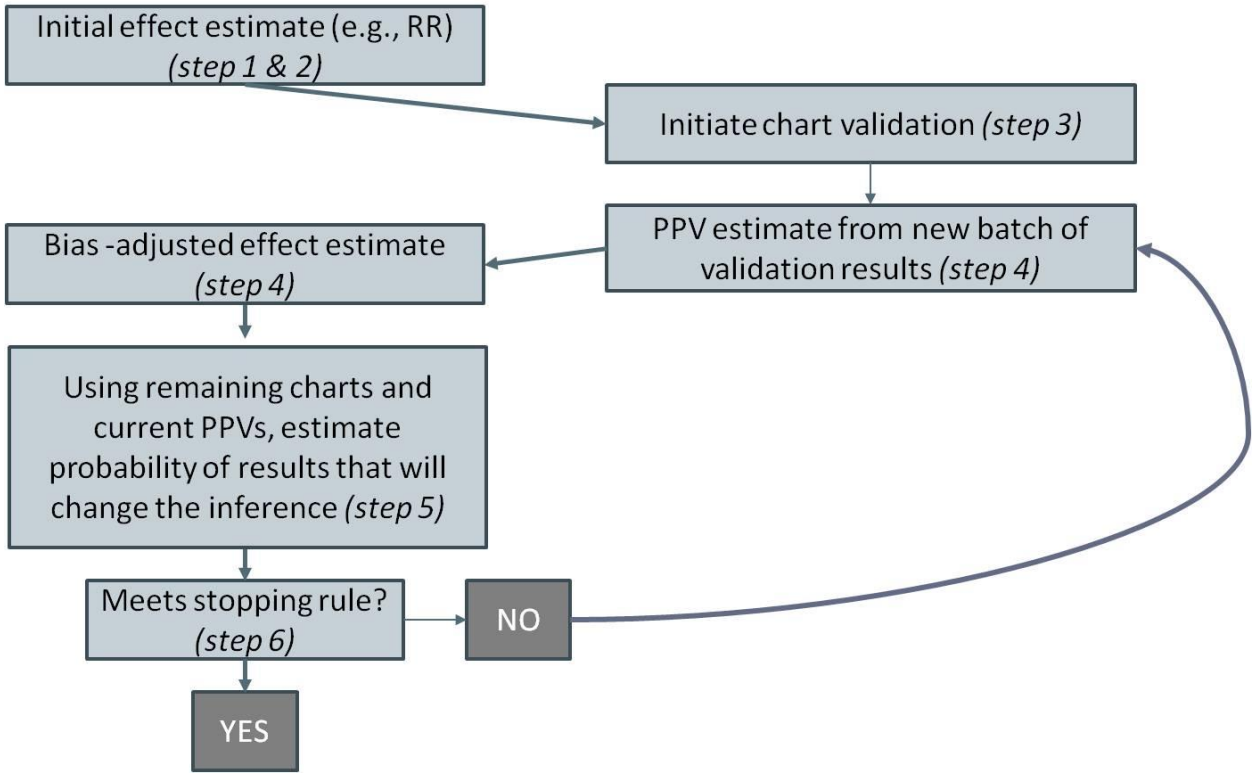
- At each batch, a known number of exposed and unexposed outcomes are yet to be validated. All possible combinations of medical chart validation results are identified.

- PBAs are then performed to determine which of the possible combinations of chart validation results would lead to adjusted results that would meet the objective of interest.

- The probability that the validation results of outstanding charts would be one of the combinations that would meet the objective of interest is calculated. This is done by using the PPVs from the cumulative validation results as priors and sampling from these distributions to determine the likelihood that a combination that would lead to adjusted results that meet the criterion of interest would be realized from the outstanding chart validation results.

Step 7 – Terminate validation process if threshold is met, otherwise continue: If the pre-specified threshold is met at a given sequential adjustment, the medical chart validation process is stopped. Otherwise, Step 5 is repeated with the results of the next batch.

Figure 3. Overview of proposed methods framework

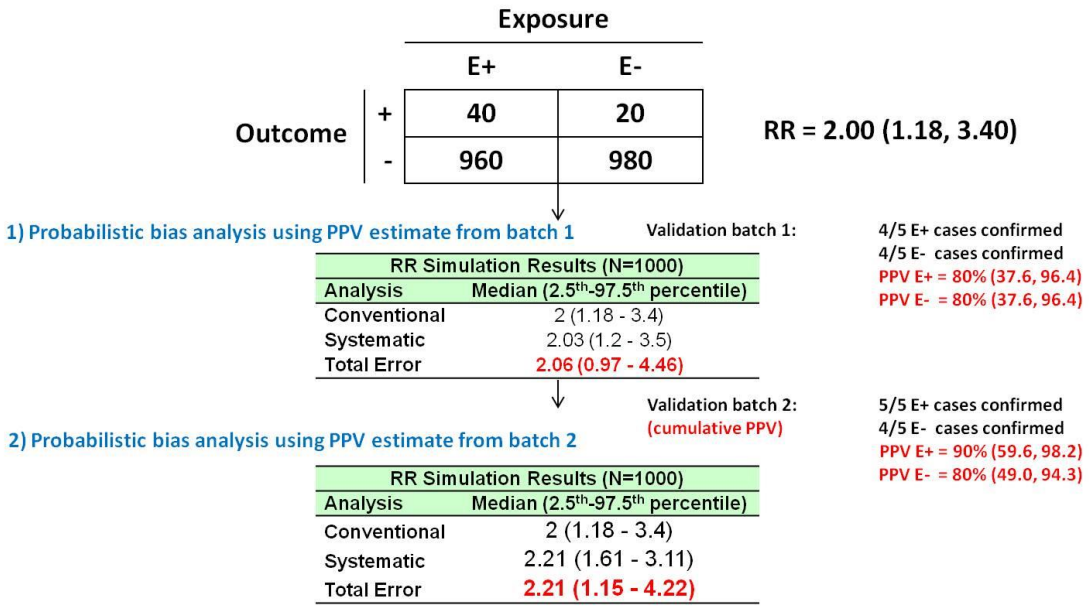


We applied the framework using the hypothetical example introduced above. The data in the 2x2 table in **Figure 4** depict the initial effect estimate from Step 2 of the framework. We assume that Step 1 was performed. Let us assume that we initiate the chart validation (Step 3), using the balanced design, and sample 20 exposed and 20 unexposed outcomes for validation. In the first batch of charts for which results became available, four out of five exposed and four out of five unexposed outcomes were confirmed as true cases. Thus, the initial PPV estimates from Step 4 were 80% with 95% CIs from 37.6% to 96.4% in both groups. Using these initial PPV estimates as input parameters (by specifying a beta distribution around them with variance based on the 95% CIs) in Step 4, and assuming an NPV of 1.00 for the outcome definition, we performed 1,000 PBA simulations and obtained a median bias-adjusted RR of 2.06 with a 95% simulation interval ranging from 0.97 to 4.46. The bias-adjusted RR incorporates uncertainty in both the PPV estimates and the original study, and has a 95% simulation interval that is wider than the 95% CI from the initial RR.

In Step 5, with the remaining 30 outcomes to be validated (15 exposed and 15 unexposed), we calculated the probability *(ζ)* of observing a bias-adjusted estimate with a 95% simulation interval including 1.00, given the observed PPVs after the first batch of validation results. We selected an arbitrary stopping rule of $\zeta < 10\%$. That is, the process of receiving additional batches and repeatedly updating the bias-adjusted RR would continue until it was determined that the probability of obtaining a bias-adjusted RR with a 95% simulation interval that included 1.00 was <10%.

With the 30 outcomes yet to be validated, there were 16 possible validation results within each exposure group (i.e., for each exposure between 0 and 15 cases could be confirmed), representing 256 (i.e., 16 x 16) possible combinations of validation results. We then conducted a PBA for each combination to determine whether that combination would lead to a bias-adjusted RR with a simulation interval that included 1.00. Each PBA used the median bias-adjusted RR and 95% simulation (i.e., 2.06 [0.97 to 4.46]) interval and the corresponding PPVs and their 95% CIs for the given combination. We identified from all 256 PBAs which combinations led to bias-adjusted RRs with simulation intervals that included 1.00. We then estimated the probability that the validation results for the outstanding outcomes would match one of these combinations. In particular, we specified independent binomial distributions with probabilities of success equal to PPV E+ and PPV E- from batch 1 and number of trials equal to the number of outcomes yet to be validated (i.e., 16 in each group). Because the PPVs are estimated with uncertainty, the actual probability of success was drawn from a beta distribution with mean equal to either PPV E+ or PPV E- point estimates and a variance based on the 95% CI of the PPV. We repeated this sampling process 1,000 times and enumerated the proportion of instances in which the number of successes in each group corresponded to one of the combinations identified above. After the first validation batch, the calculated ζ was 26.8%.

Figure 4. Results of two analysis batches from a hypothetical example



Let us assume that in the second batch, five out of five exposed outcomes were confirmed and four out of five unexposed cases were confirmed. The cumulative PPV E+ was 90% (i.e., 9 out of 10) with a 95% CI from 59.6% to 98.2%. The updated PPV E- was 80% with a 95% CI from 49.0% to 94.3%. Using the updated PPV estimates to adjust the original data yielded a median bias-adjusted RR of 2.21 (95% simulation interval, 1.15 to 4.22). After incorporating the second validation batch, the calculated ζ was 3.6%, which met the stopping rule. In other words, given the current bias-adjusted RR of 2.21 (95% simulation interval, 1.15 to 4.22), the current PPV E+ of 90% (95% CI, 59.6% to 98.2%) and PPV E- of 80% (95% CI, 49.0% to 94.3%), that there were 20 more outcomes to be validated (10 per exposure group), and the simplifying assumptions (e.g., NPV = 1.00), there was a 3.6% probability that validating the remaining outcomes would lead to a final bias-adjusted RR with a 95% simulation interval that included 1.00.

E. STAGE 4: APPLICATION OF FRAMEWORK TO AN EMPIRICAL EXAMPLE

1. Steps 1 and 2

We selected as an empirical example a previously conducted Sentinel assessment that evaluated the risk of intussusception following rotavirus vaccination.[14] This assessment examined different vaccines and doses using multiple designs. We used unadjusted person-time and outcome counts available in the final report from one particular analysis that used a cohort design and focused on one specific vaccine. We used these data to estimate a crude RR of 1.22 (95% CI, 0.83 to 1.82), which served as the preliminary RR obtained in Step 2 for the purposes of the proposed method. It is important to note that this RR is not particularly meaningful since it is not adjusted for confounding and is based on all outcomes prior to validation whereas the Sentinel assessment adjusted for confounding and only included those outcomes confirmed by medical chart validation in the primary analyses. Bias analysis methods do exist that allow adjustment for confounding and outcome misclassification simultaneously, but we focused only on outcome misclassification. This example allowed us to apply and examine the proposed method with actual empirical Sentinel data. The three DPs that contributed to this assessment permitted the Workgroup to reuse the chart validation results for the purposes of this method project.

Let us assume that two objectives of the method were specified *a priori* in Step 1. In particular, assume that we were interested in determining whether we could rule out a bias-adjusted RR with a 95% simulation interval that excluded 1.00 and that we were also interested in determining whether we could rule out a bias-adjusted RR with a 95% simulation interval that included 2.00. Let us also assume that our pre-specified decision threshold was 1% for each objective. For example, if we terminated the medical chart validation process at a given point in time, we would want to ensure that there was no more than a 1% chance that, if we had continued it, the 95% simulation interval around the bias-adjusted RR would exclude 1.00 and, separately, include 2.00.

2. Steps 3 and 4

There was a total of 176 outcomes (37 exposed and 139 unexposed) to be validated across the three DPs; 96 total outcomes in DP1; 10 in DP2; and 70 in DP3. Clinical adjudicators were blinded to vaccination history and were instructed to classify cases using Brighton Collaboration criteria.[15] Brighton Level 1 cases were considered confirmed and used to calculate PPV estimates. As medical chart validation was performed as part of the original Sentinel assessment (Step 3), we replicated the batched manner in which medical chart validation results became available to the SOC (Step 4). In what follows, we describe the application of the method as if the medical chart validation results had been batched chronologically across the three DPs and separately as if the results had been batched by DP. The former approach may be preferred when minimizing the time required for medical chart validation is the primary objective and the latter approach may be preferred when minimizing the cost of the validation is the primary objective.

Table 1 shows the medical chart validation results, across the three DPs, ordered by the date that the SOC received the validation results from the adjudicator ("final adjudication date"). For example, on February 23, 2012, results were returned for four exposed outcomes and 37 unexposed outcomes; 2 and 17 were confirmed as true cases, respectively. For simplicity, we grouped the results into four batches after results for approximately 25%, 50%, 75% and 100% of exposed outcomes had been received by the SOC. In practice, the analysis could be repeated each time any results are returned to the SOC or at any other frequency. At each batch, we calculated the cumulative PPVs and 95%

confidence intervals among exposed and unexposed outcomes. The four rows in Table 1 that are highlighted in blue indicate the four time points at which adjustments were made in Table 3.

Table 1. Cumulative chart validation results over time, across three Data Partners

Final adjudication date	Exposed outcomes		Unexposed outcomes	
	Confirmed cases	PPV E+ (95% CI)	Confirmed cases	PPV E- (95% CI)
23 February 2012	2/4		17/37	
06 March 2012	2/4		17/38	
16 March 2012	2/4		21/47	
25 March 2012	2/5		21/47	
09 April 2012	3/7		25/55	
03 May 2012	**4/11**	**36.3 (15.2, 64.6)**	**30/67**	**44.8 (33.2, 56.8)**
14 May 2012	5/12		38/79	
29 May 2012	6/13		41/86	
04 June 2012	6/14		41/86	
22 June 2012	6/15		43/90	
22 June 2012	6/15		45/94	
1 August 2012	**10/19**	**52.6 (31.7, 72.7)**	**48/99**	**48.5 (38.9, 58.2)**
22 August 2012	10/20		48/99	
30 August 2012	11/23		52/106	
25 September 2012	12/24		58/117	
14 November 2012	12/24		59/118	
14 November 2012	13/26		61/120	
21 November 2012	**16/29**	**55.2 (37.5, 71.6)**	**64/123**	**52.0 (43.3, 60.7)**
28 November 2012	18/32		67/129	
30 November 2012	18/32		67/130	
30 November 2012	18/36		72/138	
10 December 2012	**18/37**	**48.6 (33.4, 64.1)**	**72/139**	**51.8 (43.6, 59.9)**

Table 2 shows the medical chart validation results batched by DP with cumulative PPV estimates and 95% CIs updated with the addition of each DP.

Table 2. Cumulative chart validation results across three Data Partners

Data Partners included	Exposed outcomes		Unexposed outcomes	
	Confirmed cases	PPV E+ (95% CI)	Confirmed cases	PPV E- (95% CI)
DP1	6/11	54.5 (28.0, 78.7)	39/85	45.8 (35.6, 56.4)
DP1 + DP2	6/14	42.9 (19.6, 68.9)	44/92	47.8 (37.8, 58.0)
DP1 + DP2 + DP3	18/37	48.6 (33.4, 64.1)	72/139	51.8 (43.6, 59.9)

3. Steps 5, 6, and 7

Batching by adjudication date: Table 3 shows the results of the sequential bias-adjustment and the calculation of ζ for both objectives when medical chart validation results were batched over time by adjudication date. We conducted four sequential PBA adjustments using the cumulative PPV estimates at each of the four specified batches. Based on the PPV estimates obtained at the first batch (i.e., 36.3% [95% CI, 15.2% to 64.6%] for PPV E+ and 44.8% [95% CI, 33.2% to 56.8%] for PPV E-), the first bias-adjusted median RR was 0.98 with a 95% simulation interval ranging from 0.46 to 1.79. At this point, the validation results of 98 outcomes were still outstanding. Given this, we calculated the probability (ζ) of obtaining a bias-adjusted median RR with a 95% simulation interval to be 2.31% if the remaining 98 outcomes were validated and using the observed PPV estimates and their 95% CIs as prior distributions for those remaining validation results. After the third sequential adjustment, we obtained a median bias-adjusted RR of 1.29 (95% CI, 0.82 to 2.03), and a corresponding ζ *of* 0.55%, which met our arbitrary stopping rule of 1%.

We also calculated the probability (ζ) of observing a bias-adjusted RR with a simulation interval including 2.00. After the first batch of validation results, we calculated a 16.60% chance that continuing the validation process would result in a median bias-adjusted RR with a 95% simulation interval that included 2.00. This probability increased with the second batch and was 53.60% after the third batch, which did not meet the stopping threshold. After incorporating all of the results of the validation process, the final bias-adjusted median RR was 1.14 with a 95% simulation interval (0.73 to 1.79) that did exclude 2.00.

Table 3. Sequential bias adjustment by adjudication date

	Bias-adjusted RR, median (2.5^{th}-97.5^{th} percentile)	Probability (ζ) of 95% simulation interval excluding 1.00 with outstanding cases	Probability (ζ) of 95% simulation interval including 2.00 with outstanding cases
Adjustment 1	0.98 (0.46, 1.79)	2.31%	16.60%
Adjustment 2	1.32 (0.80, 2.13)	10.36%	58.34%
Adjustment 3	1.29 (0.82, 2.03)	0.55%	53.60%
Adjustment 4	1.14 (0.73, 1.79)	(no charts remaining)	(no charts remaining)

Batching by DP: Table 4 displays the results of the sequential bias adjustment and the calculation of *ζ* when the medical chart validation results were cumulatively batched by DP and for both objectives. Using only validation results from DP1 (n = 96 outcomes), we obtained a median bias-adjusted RR of 1.44 with a 95% simulation interval ranging from 0.81 to 2.41. With 80 outcomes yet to be validated from DP2 and DP3 and given the validation results from DP1, the calculated probability of obtaining a bias-adjusted median RR with a 95% simulation interval that excluded 1.00 was 33.20%. When validation results from DP2 (n = 10 outcomes) were incorporated, the median bias-adjusted RR was 1.09 (95% simulation interval, 0.59 to 1.97) and the *ζ* was 2.46%, which did not meet the stopping rule, so results from DP3 were needed.

For the objective of ruling out a bias-adjusted RR with a 95% simulation interval that included 2.00, we obtained ζ = 73.48% with validation results from only DP1 and ζ = 21.13% adding validation results from DP2. As such, it was determined that validation results from DP3 would be needed.

Table 4. Sequential bias adjustment by Data Partner

	Bias-adjusted RR, median (2.5th-97.5th percentile)	**Probability *(ζ)* of 95% simulation interval excluding 1.00 with outstanding cases**	**Probability *(ζ)* of 95% simulation interval including 2.00 with outstanding cases**
DP1	1.44 (0.81, 2.41)	33.20%	73.48%
DP1 + DP2	1.09 (0.59, 1.97)	2.46%	21.13%
DP1 + DP2 + DP3	1.14 (0.73, 1.79)	(no charts remaining)	(no charts remaining)

II. DISCUSSION

We propose an approach to sequential application of QBA methods to adjust for bias due to outcome misclassification that quantifies the likelihood that continuing the medical chart validation process to inform the bias adjustment will meet a certain objective. The method has the potential to reduce the time and cost associated with medical chart validation in Sentinel assessments and other observational studies using healthcare claims data, where expedited evidence generation is important and resources are limited.

In the empirical analyses in which we batched chart validation results by adjudication date and calculated the probability of 95% simulation interval excluding 1.00, we met the arbitrary stopping rule of a probability of <1% after including results from three of four batches. If the validation process had been stopped, the method would have resulted in up to an approximately 25% reduction in cost, since each batch contained roughly 25% of total charts reviewed. This assumes a fixed cost for each chart and assumes that costs are incurred only for those for which SOC received results. The method would have suggested stopping the validation process almost three weeks before SOC received the final chart validation results. These potential monetary and time savings would need to be balanced with the time and effort required to conduct the sequential analysis.

In proposing and applying the method, we made a number of simplifying assumptions for both the PBA and the calculation of ζ in both the hypothetical and empirical examples. In conducting the PBAs, we

assumed that NPV among both exposed and unexposed outcomes was 100%, which implies perfect sensitivity. We acknowledge that a sensitivity of 1.00 is likely not tenable for most claims-based outcomes definitions. However, in the setting of rare outcomes, NPV tends to be very close to 1.00, even with relatively low sensitivity. Nevertheless, minor variations in NPV estimates can lead to large changes in bias-adjusted estimates. The PBA framework can not only accommodate other values of NPV, but can also account for NPV distributions, which reflect uncertainty in the true NPV, much like the distributions used for PPV. Varying NPV, such as by sampling from a distribution, will further increase uncertainty. Further parameterizing NPV within the proposed methods framework should be explored in future work.

We focused on count data for simplicity. As such, we considered classification of binary outcomes and did not worry about potential misclassification of the timing of the outcomes. QBA methods can also be used for time-to-event outcomes. Also, we considered only two possible levels of adjudication in our analysis – namely, whether or not outcomes were true confirmed as true cases. Even with a medical chart, it is not always possible to make clear determinations about whether an outcome is or is not a true case. Medical chart validations often include other categories, such as possible or probably, to indicate uncertainty in the adjudication. In theory, this uncertainty could be incorporated into the QBA by, for example, assigning probabilities to outcomes in each adjudication category.

Possible correlation between PPV E+ and PPV E- estimates was not considered, but could be addressed within the methods framework. In addition, as with any semi-Bayesian method, the approach is sensitive to the choice of distributions and priors. We recommend sensitivity analyses that vary these assumptions when using the method. Future work could focus on expanding the method to a full Bayesian approach, which would require a prior on the estimand. Thus far, the approach has focused on adjusting for outcome misclassification, but it could also be used in the context of multiple bias adjustment in which multiple sources of potential bias are simultaneously addressed.

In our applications of the approach, we assumed that cumulative observed PPVs were reasonable estimates of PPVs for outcomes that had not yet been validated. When charts are batched based on the final adjudication date, this assumption is tenable to the extent that the order in which the adjudicator receives and reviews charts does not depend on the validation results. We cannot rule out the possibility that practice sites with fewer resources may be associated with higher misclassification rates and may take longer to send charts to the adjudicator. In batching by DP, we assumed that the observed PPV estimates from one DP were reasonable priors for the PPV estimates in other DPs. In practice, PPVs may vary across DPs if, for example, the specificity of the outcome definition is the same but the underlying true incidence of the outcome varies across DPs. Nevertheless, while we used empirical priors for the calculation of ζ, priors that incorporate different beliefs could also be used.

III. CONCLUSION

The proposed method is flexible with respect to the objective of interest (e.g., ruling out a bias-adjusted RR with a 95% simulation interval that excludes 1.00 or with a 95% simulation interval that includes 2.00) and the chosen stopping threshold (e.g., <1%). We do not prescribe the use of any particular threshold, as it is likely that the objective and stopping threshold will be tailored to a given assessment based on the hypothesized magnitude of effect of the exposure, the public health impact of the outcome under study, the availability of prior external validation information, and other study-specific considerations. While the current method provides clear quantitative criteria for stopping chart validation early, whether the implementation of the methods translates into actual cost and time savings should be evaluated in future work.

IV. ACKNOWLEDGEMENTS

The Workgroup thanks Carolyn Balsbaugh, MPH, and Katherine Yih, PhD, MPH, for their support in understanding the medical chart validation process within Sentinel and for collating the medical chart validation results. We also thank the three Data Partners (including Aetna, Inc. and Humana) who permitted use of the chart validation results for this project.

V. REFERENCES

1. Gagne JJ, Fireman B, Ryan PB, et al. Design considerations in an active medical product safety monitoring system. *Pharmacoepidemiol Drug Saf*. 2012;21 Suppl 1:32-40.

2. Carnahan RM. Mini-Sentinel's systematic reviews of validated methods for identifying health outcomes using administrative data: summary of findings and suggestions for future research. *Pharmacoepidemiol Drug Saf*. 2012;21 Suppl 1:90-99.

3. Lash T, Fox M, Fink A. *Applying Quantitative Bias Analysis to Epidemiological Data*.

4. Strom BL, Carson JL, Halpern AC, et al. Using a claims database to investigate drug-induced Stevens-Johnson syndrome. *Stat Med*. 1991;10(4):565-576.

5. Eng PM, Seeger JD, Loughlin J, Clifford CR, Mentor S, Walker AM. Supplementary data collection with case-cohort analysis to address potential confounding in a cohort study of thromboembolism in oral contraceptive initiators matched on claims-based propensity scores. *Pharmacoepidemiol Drug Saf*. 2008;17(3):297-305.

6. Andrade SE, Gurwitz JH, Chan KA, et al. Validation of diagnoses of peptic ulcers and bleeding from administrative databases: a multi-health maintenance organization study. *J Clin Epidemiol*. 2002;55(3):310-313.

7. Schneeweiss S, Robicsek A, Scranton R, Zuckerman D, Solomon DH. Veteran's affairs hospital discharge databases coded serious bacterial infections accurately. *J Clin Epidemiol*. 2007;60(4):397-409.

8. Psaty BM, Kuller LH, Bild D, et al. Methods of assessing prevalent cardiovascular disease in the Cardiovascular Health Study. *Ann Epidemiol*. 1995;5(4):270-277.

9. Leveille SG, Gray S, Black DJ, et al. A new method for identifying antibiotic-treated infections using automated pharmacy records. *J Clin Epidemiol*. 2000;53(10):1069-1075.

10. Sidney S, Sorel M, Quesenberry CP, DeLuise C, Lanes S, Eisner MD. COPD and incident cardiovascular disease hospitalizations and mortality: Kaiser Permanente Medical Care Program. *Chest*. 2005;128(4):2068-2075.

11. Brown NJ, Ray WA, Snowden M, Griffin MR. Black Americans have an increased rate of angiotensin converting enzyme inhibitor-associated angioedema. *Clin Pharmacol Ther*. 1996;60(1):8-13.

12. Holcroft CA, Spiegelman D. Design of validation studies for estimating the odds ratio of exposure-disease relationships when exposure is misclassified. *Biometrics*. 1999;55(4):1193-1201.

13. Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. *Int J Epidemiol*. 2014;43(6):1969-1985.

14. Yih WK, Lieu TA, Kulldorff M, et al. Intussusception risk after rotavirus vaccination in U.S. infants. *N Engl J Med*. 2014;370(6):503-512.

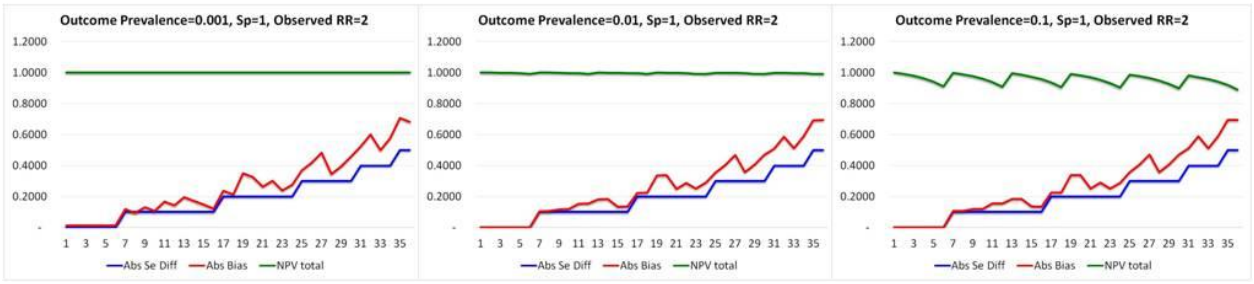
15. Bines JE, Kohl KS, Forster J, et al. Acute intussusception in infants and children as an adverse event following immunization: case definition and guidelines of data collection, analysis, and presentation. *Vaccine*. 2004;22(5-6):569-574.

VI. APPENDIX

We constructed three scenarios and conducted simulations in each to understand the impact of imperfect negative predictive value (NPV). In all three scenarios, we simulated data such that the observed association between the exposure and outcome was a risk ratio (RR) of 2.0. Across each scenario we simulated all 36 possible combinations of the following values for sensitivity in the exposed group (Se1) and sensitivity in the unexposed group (Se0): 0.5, 0.6, 0.7, 0.8, 0.9, and 1.0. In each scenario we varied the specificity (Sp): Sp = 1.00 in Scenario 1, Sp = 0.999 in Scenario 2, and Sp = 0.995 in Scenario 3. The same specificity was used for both exposure groups. We also varied the true outcome prevalence within each scenario, with values of 0.001, 0.01, and 0.1 in Scenario 1 and values of 0.01 and 0.1 in Scenarios 2 and 3. For each outcome prevalence in each scenario, we plotted the NPV and degree of differential sensitivity, defined as |Se1 – Se0| along with absolute bias, which we calculated as |ln(observed RR) – ln(bias-adjusted RR)|.

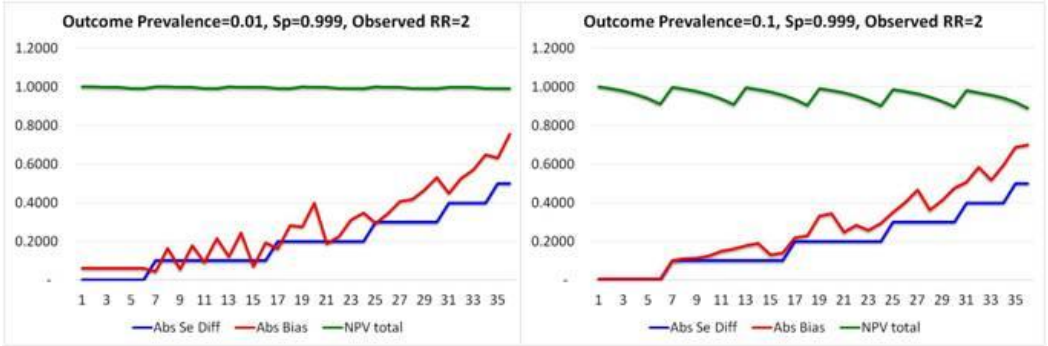
Scenario 1 results

In Scenario 1, when specificity was 1.000, results across the three values of outcome prevalence were virtually identical. Absolute bias tracked with changes in absolute difference in sensitivity between exposure groups. At each "step" representing the absolute difference in sensitivity between exposure groups, bias fluctuated as a function of the sensitivity values, but did not appear to correlate with NPV values. For example, with outcome prevalence of 0.001, in combinations where the absolute difference in sensitivity was 0.10 (e.g., combinations such as 1.0 vs. 0.9 and 0.5 vs. 0.5), bias ranged from 0.0931 (combination 8) to 0.1946 (combination 13). The NPVs in these combinations were 0.9999 and 1.0000, respectively.



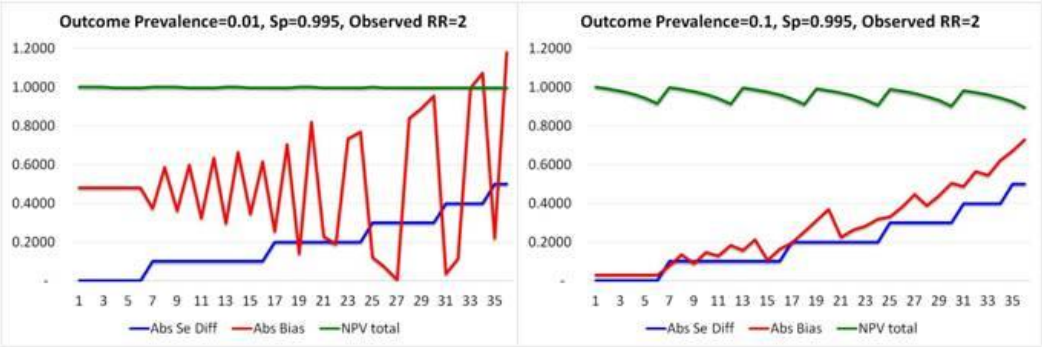
Scenario 2 results

Similar results were observed in scenario 2, where the specificity was high, but not perfect (i.e., 0.999).



Scenario 3 results

In Scenario 3, when the outcome prevalence was 0.01, bias was large across most combinations of sensitivity values, even when sensitivity was perfect in both exposure groups (i.e. combination 1). Absolute bias varied widely across combinations with the same absolute difference in sensitivity, but the degree of variation generally increased with increasing difference in absolute sensitivity. Again, absolute bias was driven by the actual values of sensitivity rather than the overall NPV. For example, when the absolute difference in sensitivity was 0.20, combination 20 yielded an NPV of 0.9990 and absolute bias of 0.8182 while combination 21 yielded a low NPV of 0.9985 and also a lower absolute bias of 0.2259. With a more common outcome (i.e., prevalence = 0.1), Scenario 3 results were generally similar to those of Scenarios 1 and 2, where bias was low with non-differential sensitivity and increased with increasing differences in sensitivity between exposure groups.



Overall, these simulation results suggest that, when specificity and outcome prevalence are low, bias can be large even with perfect sensitivity and NPV. They also suggested that, at a given specificity and outcome prevalence, differences in sensitivity between exposure groups can have a large impact on absolute bias even with very high NPV values. Based on these findings, the Workgroup recommends that, while assumptions about high NPV may be tenable, NPV values should be varied within the proposed method to reflect potential differences in sensitivity between exposure groups.