OPEN

Adaptive Validation Design

A Bayesian Approach to Validation Substudy Design With Prospective Data Collection

Lindsay J. Collin,[a] *Richard F. MacLehose,*[b] *Thomas P. Ahern,*[c] *Rebecca Nash,*[a] *Darios Getahun,*[d] *Douglas Roblin,*[e] *Michael J. Silverberg,*[f] *Michael Goodman,*[a] *and Timothy L. Lash*[a]

Abstract: An internal validation substudy compares an imperfect measurement of a variable with a gold-standard measurement in a subset of the study population. Validation data permit calculation of a bias-adjusted estimate, which has the same expected value as the association that would have been observed had the gold-standard measurement been available for the entire study population. Existing guidance on optimal sampling for validation substudies assumes complete enrollment and follow-up of the target cohort. No guidance exists for validation substudy design while cohort data are actively being collected. In this article, we use the framework of Bayesian monitoring methods to develop an adaptive approach to validation study design. This method monitors whether sufficient validation data have been collected to meet predefined criteria for estimation of the positive and negative predictive values. We demonstrate the utility of this method using the Study of Transition, Outcomes and Gender—a cohort study of transgender and gender nonconforming people. We demonstrate the method's ability to determine efficacy (when sufficient validation data have accumulated to obtain estimates of the predictive values that fall above a threshold value) and futility (when sufficient validation data have accumulated to conclude the mismeasured variable is an untenable substitute for the gold-standard measurement). This proposed method can be applied within the context of any parent epidemiologic study design and modified to meet alternative criteria given specific study or validation study objectives. Our method provides a novel approach to effective and efficient estimation of classification parameters as validation data accrue.

Keywords: Bayesian methods; Validation study design

(*Epidemiology* 2020;31: 509–516)

Validation studies compare an imperfect measurement of a variable with a gold-standard measurement for the same variable, usually in a subset of the parent study population. The gold-standard measurement is often too expensive or too difficult to obtain for all participants. Validation data allow researchers to quantify the degree to which the imperfect measurement approximates the gold standard, which can be used to inform a bias-adjusted estimate of association in etiologic research or occurrence in surveillance research. The importance of validation studies is widely recognized in epidemiologic research and is encouraged by funding agencies, which expect supported research to be rigorous and reproducible.[1] Historically, there has been a lack of validation data available, and some journals have encouraged publication of validation study results as independent articles.[2,3] Little attention, however, has been given to effective and efficient designs for validation substudies.

Existing guidance on optimal sampling of participants for a validation substudy, such as the balanced design or a simple random sample, pertains only to scenarios in which

Submitted September 4, 2019; accepted May 4, 2020.

From the [a]Department of Epidemiology, Emory University, Atlanta, GA; [b]Division of Epidemiology and Community Health, University of Minnesota, Minneapolis, MN; [c]Department of Surgery, The Robert Larner, M.D. College of Medicine at The University of Vermont, Burlington, VT; [d]Department of Research and Evaluation, Kaiser Permanente Southern California, Pasadena, CA; [e]Mid-Atlantic Permanente Research Institute, Kaiser Permanente Mid-Atlantic States, Rockville, MD; and [f]Division of Research, Kaiser Permanente Northern California, Oakland, CA.

Supported in part by the US National Cancer Institute (F31CA239566 awarded to L.J.C., and R01CA234538 awarded to T.L.L.) and the US National Library of Medicine (R01LM013049 awarded to T.L.L.). T.P.A. was supported by an award from the US National Institute of General Medical Sciences (P20 GM103644). STRONG cohort data were collected with support from Contract AD-12-11-4532 from the Patient Centered Outcome Research Institute and by the National Institute of Child Health and Human Development R21HD076387 awarded to M.G.

The authors report no conflicts of interest.

SDC Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com).

Data Access: Due to patient confidentiality, data are only available upon Institutional Review Board approval from the research institution in collaboration with Dr. Michael Goodman and the STRONG research team. Example code used to perform the adaptive validation studies is available from GitHub (https://github.com/lcolli5/Adaptive-Validation).

Correspondence: Lindsay J. Collin, Department of Epidemiology, Emory University, 1518 Clifton Rd NE Atlanta, GA 30322. E-mail: lindsay.jane.collin@emory.edu.

ISSN: 1044-3983/20/3104-0509
DOI: 10.1097/EDE.0000000000001209

the complete parent study population has been enrolled and follow-up has been completed.[3,4] Previous work on designing validation substudies has focused on setting a sample size required to estimate the bias parameters with a prespecified degree of precision, given fixed resources available to support the validation study.[5–8] These methods assume that the complete parent study population is available for inclusion in the validation study. Conversely, given the cost of implementing a validation study, researchers may want to know at what point sufficient validation data have been collected to meet the objectives of validation. Prospective monitoring of validation data as they accrue allows researchers to determine when sufficient validation data have been collected to obtain classification parameters that meet stopping criteria. For example, validation data may be collected to decide on which side of a threshold value the classification parameters fall. The threshold can be used to decide whether the classification is so poor that the gold-standard measure is needed for everyone, or that classification is so good that further validation is unnecessary. Further, an adaptive validation study could be used to assure sufficient precision of a bias-adjusted estimate of effect or occurrence, or to detect a time trend in the classification parameters. We are unaware of any published guidelines for the design and implementation of validation substudies in which validation data are longitudinally collected, let alone with prespecified stopping rules as opposed to a fixed sample size. This scenario can occur within the context of an internal validation substudy with any parent epidemiologic study design, whether the parent study is ongoing or completed.

Bayesian monitoring techniques have been used in clinical trials to estimate and monitor treatment response over time and to adapt the study design as data accrue, either by stopping the trial early or by modifying treatment allocation probabilities.[9] Herein, we extend this conceptual framework for the development of an adaptive approach to validation study design that informs collection of sufficient validation data to estimate the positive predictive value (PPV) and negative predictive value (NPV) for classification of a binary measure, given a predefined stopping threshold value. This strategy can be viewed as a sequential Bayesian analysis, in which the distributions of the PPV and NPV are estimated at specified intervals while the validation data accrue. For instance, one might choose to estimate the PPV and NPV each week, after every 10 participants are validated, or after each new validated measurement during the validation study. At each time point, one uses the newly collected validation data to update the estimates of PPV and NPV, to compare the results against the stopping criteria, and to decide whether to collect additional validation data. We demonstrate the utility of an adaptive validation design under 2 scenarios: (1) to determine efficacy, which occurs when sufficient validation data accumulate to yield estimates of PPV and NPV that fall above a threshold value and (2) to determine when collecting further validation data is futile, which occurs when sufficient validation data accumulate to show that the mismeasured variable lacks face validity and the gold standard may need to be measured for the complete parent study population.

METHODS

Parent Study Population

As an exemplar, we use the Study of Transition, Outcomes and Gender (STRONG) cohort.[10] The STRONG transgender and gender nonconforming cohort was established to understand long-term effects of hormone therapy and surgery on gender dysphoria and mental health, as well as cardiovascular disease, metabolic or endocrine disorders, and cancer incidence following hormone therapy or gender-affirming surgery. Cohort members were identified using International Classification of Diseases, 9th Revision (ICD-9) codes and keywords related to gender dysphoria from Kaiser Permanente health plans in Georgia, Northern California, and Southern California. This study was approved through Emory University Institutional Review Board (IRB) (#0006742). Participant consent was not required as the study used de-identified data obtained from the Kaiser Permanente sites. Each Kaiser Permanente site received its own IRB approval.

The index date corresponds to entry into the cohort and is the first date with a recorded ICD-9 code or keyword reflecting transgender and gender nonconforming status between 2006 and 2014. Demographic data collected from the electronic medical record included the patient's gender, but it was unclear whether a given person's gender in the medical record corresponded to their gender identity or to their sex assigned at birth. This misclassification precluded accurate assignment of cohort members to transfeminine or transmasculine status based on their electronic medical record gender code. Therefore, cohort members' archived medical records—the gold standard in this study—were reviewed to determine sex assigned at birth. All persons included in this aspect of the study are transgender, so sex assigned at birth determines transfeminine or transmasculine status. Medical records were reviewed by keyword search in selected text strings to identify additional anatomy- or therapy-related terms that would indicate sex assigned at birth. The STRONG cohort was divided into 2 subcohorts, the first including youths (1,331 persons <18 years of age) and the second including adults (4,725 persons ≥18 years of age). Persons under 18 years of age with a recorded ICD-9 code or keyword reflecting transgender or gender nonconforming status are less likely to have initiated hormone therapy or gender-affirmation surgery than adults, so the electronic medical record's gender code is more likely to reflect their sex assigned at birth than their gender identity.

Exposure Validation

The exposure of interest was sex assigned at birth, which must be known to ascertain transfeminine/transmasculine status. Sex assigned at birth was determined from the concurrent electronic medical record demographic data and known to be misclassified because it could either represent sex assigned at birth or concurrent gender. The demographic gender variable was validated for members of the youth subcohort who were ≥18 years of age as of 1 January 2015 (n = 535; 40% of the youth subcohort). For demonstration of the method, we only included these 535 youth subcohort members. Validation of the electronic medical record gender code was completed for all STRONG adult subcohort members. The validation that was completed in each subcohort allowed for comparison of classification parameter estimates between our adaptive validation approach and the complete validation approach originally used by the STRONG study. Although the method we propose is general enough to incorporate a variety of possible bias parameters, we focus on the PPV—defined as the probability that sex assigned at birth was male among those whose electronic medical record gender code was male—and NPV—defined as the probability that sex assigned at birth was female among those whose electronic medical record gender code was female. Our assignments of the labels "male" and "female" to "positive" and "negative" predictive values, respectively, were made at random.

Analytic Strategy

We used gender as recorded on the STRONG cohort's electronic medical record and the validation data on sex assigned at birth to calculate the classification parameters (PPV and NPV). Although the STRONG cohort had already completed enrollment, validation, and follow-up, we conducted the validation substudy as though it was contemporaneous with parent study data collection for the original cohort, and compared the estimates from our method with those obtained from analyzing the complete validation data. We ordered youth and adult subcohort members chronologically by their index date, which allowed selection of participants as data would have accrued over time. We applied our Bayesian adaptive validation methods to identify the sample size necessary for estimates of PPV and NPV to meet stopping criteria, which we define below, as this time passed.

Once cohort members were ordered by index date, we used an iterative beta-binomial Bayesian model to update the PPV and NPV at regular follow-up intervals.[11] We assume, before seeing any validation data, that all values of PPV and NPV are equally likely, due to the lack of prior information on recording gender identity in electronic medical records. In settings where a literature exists, it may be used to inform this prior distribution by incorporating information from previous validation studies. Because the same updating process will be completed independently for PPV and NPV within the youth or adult subcohorts, we refer to both parameters as θ :

$$prior: \theta \sim beta(1,\ 1)$$

This beta distribution is identical to a uniform distribution, with all predictive values having equal probability. The data in this validation study correspond to whether the *i*th individual's gold-standard sex assigned at birth, y_i, matched the observed electronic medical record gender code. That is, if a person's observed gender code was male, did the medical record indicate the person's sex assigned at birth was male (for PPV), or if the person's observed gender code was female, did the medical record indicate the person's sex assigned at birth was female (for NPV). The likelihood contributed by the *i*th individual in the validation substudy is:

$$\boldsymbol{data}: \boldsymbol{y_i} \mid \theta \sim \boldsymbol{Bernoulli}(\theta)$$

After the first individual's validation data have been collected, the likelihood and prior can be combined *via* Bayes' theorem, and the posterior distribution of the bias parameter was calculated as follows:

$$\boldsymbol{Posterior}: \ \theta \mid \boldsymbol{y}_1 \sim \ \boldsymbol{Beta}\,(\boldsymbol{a}_1, \boldsymbol{b}_1)$$

Where $a_1 = 1 + y_1$ and $b_1 = 2 - y_1$. The mean of this distribution can be used as an estimate of the bias parameter: $E(\theta) = \frac{a_1}{(a_1 + b_1)}$. One could choose to use the median or mode of the distribution instead, and credible intervals can be defined using percentiles of the distribution. After the first observation has been accumulated, the posterior distribution becomes the prior distribution that is updated by the second observation, and this process repeats for each observation, or block of observations, collected:

$$Time1: p(\theta \mid y_1) \propto p(y_1 \mid \theta) \times p(\theta)$$

$$Time2: p(\theta \mid y_2, y_1) \propto p(y_2 \mid \theta) \times p(\theta \mid y_1)$$

$$Time\,j: p(\theta \mid y_j, \ldots y_2, y_1) \propto p(y_j \mid \theta) \times p(\theta \mid y_{j-1}, \ldots y_2, y_1)$$

After the (*j*th) individual is validated, the posterior is:

$$\boldsymbol{Posterior}: \ \theta \mid \boldsymbol{y} \sim \ \beta(1 + \sum_{\boldsymbol{i}=1}^{\boldsymbol{J}} \boldsymbol{y_i}, 1 + \boldsymbol{J} - \sum_{\boldsymbol{i}=1}^{\boldsymbol{J}} \boldsymbol{y_i})$$

As above, the mean, median, mode, and credible intervals can be calculated from this distribution to give estimates of PPV or NPV at any point during the accumulation of validation data. This updating procedure is conducted separately for PPV and NPV estimates.

Although it is possible to update PPV and NPV after each individual is validated, it may be more practical or realistic to update in blocks. For example, blocks of participants may be validated by medical record review each day, week, or over other timeframes. To establish intervals in which validation data would be updated, we evaluated the design under several scenarios. First, when each category of the misclassified exposure table accrued a minimum cell count of 5 (mimicking a balanced validation design),[2] we randomly sampled 5 individuals without replacement from each exposure category (5-per-cell validation) to validate gender assigned at birth and then update the PPV and NPV estimates using the formula above. This validation and updating process was done repeatedly while the simulated parent study was being conducted. Second, when each category of the exposure table accrued a minimum cell count of 10 (also mimicking a balanced design), we randomly sampled 10 individuals without replacement from each exposure category (10-per-cell validation) to validate and estimate the PPVs and NPVs. Third, we validated each cohort member as they entered the study to mimic a design in which everyone gets validated. Each of these 3 validation schemes was continued until a stopping criterion had been reached, at which point the validation study was terminated.

We considered a stopping criterion based on the magnitude of the bias parameters. The validation efficacy criterion was met if the lower 95% credible bound (2.5th percentile) for the PPV and the lower 95% credible bound for NPV were >0.60, at which point validation would cease. We likewise considered validation to be futile if the upper 95% credible bound (97.5th percentile) fell below 0.60 for either one of the predictive values. Selection of 0.60 as the threshold was informed by topic content; if the classification method of gender in the electronic medical record (EMR) demographic data was not much better than a coin flip, which we operationalized as 60%, then the classification method has poor validity and the gold-standard measurement for the entire cohort would be necessary. The choice of efficacy or futility criteria for thresholds can be established by similar content-specific knowledge.

We evaluated whether the adaptive validation design outperformed other prospective validation designs by comparing our approach with a simple prospective validation design in which exposure was validated in the first 200 cohort enrollees.[4,12]

RESULTS

The STRONG youth subcohort included 535 persons with validation data, of whom 252 (47%) were classified as male in the EMR and 283 (53%) were classified as female. Enrollment into the study occurred between 2006 and 2014, with an overall study period of enrollment and follow-up of 10.5 years. The youth subcohort PPV calculated from complete validation at the end of the study enrolment period was 0.81 (95% confidence interval [CI] = 0.76, 0.86) and the NPV was 0.92 (95% CI = 0.88, 0.95) (Table 1). The STRONG adult subcohort consisted of 4,725 persons with validation data, of whom 2,409 (51%) were classified as male in the EMR and 2,316 (49%) were classified as female. The period of enrollment and follow-up lasted 10.7 years. In the adult subcohort, the overall PPV based on complete validation at the end of the study enrollment period was 0.62 (95% CI = 0.60, 0.64) and NPV was 0.49 (95% CI = 0.47, 0.51) (Table 2).

Validation Efforts Effective

In the STRONG youth subcohort, where classification was known to be accurate based on the complete data, the adaptive validation design identified when to cease validation efforts (Figure 1). In the 5-per-cell validation approach, in which 10 individuals were included in the validation set at each discrete time interval, the adaptive validation approach reached the threshold of the lower 2.5th percentile credible interval above 0.60 after collecting validation data for 120 individuals, 711 days into the study follow-up, and once 133 individuals had been enrolled. The final estimated PPV and NPV classification parameters were 0.71 (95% CI = 0.60, 0.81) and 0.87 (95% CI = 0.78, 0.94), respectively. When the 10-per-cell validation approach was applied, in which 20 individuals were included in the validation at each time interval, the adaptive validation approach similarly reached the threshold of the lower 2.5th percentile credible interval above 0.60 after collecting validation data for 120 individuals, 711 days into the study, once 133 individuals had been enrolled. The final PPV classification estimate was 0.71 (95% CI = 0.60, 0.81) and NPV was 0.87 (95% CI = 0.78, 0.94). In the single-person validation approach, the stopping rule reached 627 days into the study period with 101 individuals enrolled, after 101 individuals had their exposure status validated with PPV estimate of 0.72 (95% CI = 0.60, 0.83) and NPV of 0.85 (95% CI = 0.74, 0.94) (Figure 2).

Validation Efforts Futile

In the STRONG adult subcohort, both the PPV and the NPV quickly fell below the threshold (Figure 1), indicating futility. In the 5-per-cell validation approach, the NPV fell below the futility threshold in the second time interval,

TABLE 1. Estimates of the Classification Parameters From the 3 Approaches of the Adaptive Validation Design Among the STRONG Youth Cohort, With Comparison to the Overall Estimates From the Full Validation Cohort and Conventional Methods

Method	PPV	NPV	Number Validated	Days Into Study Period
Overall	0.81 (0.76, 0.86)	0.92 (0.88, 0.95)	535	2,897
Adaptive validation (5-per-cell)	0.71 (0.60, 0.81)	0.87 (0.78, 0.94)	120	711
Adaptive validation (10-per-cell)	0.71 (0.60, 0.81)	0.87 (0.78, 0.94)	120	711
Adaptive validation (single)	0.72 (0.60, 0.83)	0.85 (0.74, 0.94)	105	627
First 200	0.80 (0.71, 0.87)	0.92 (0.84, 0.96)	200	1,757

TABLE 2. Estimates of the Classification Parameters From the 2 Approaches of the Adaptive Validation Design Among the STRONG Adult Cohort, With Comparison to the Overall Estimates From the Full Validation Cohort and Conventional Methods

Method	PPV	NPV	Number Validated	Days Into Study Period
Overall	0.62 (0.60, 0.64)	0.49 (0.47, 0.51)	4,725	3,921
Adaptive validation (5-per-cell)	0.42 (0.17, 0.69)	0.25 (0.06, 0.52)	20	3
Adaptive validation (10-per-cell)	0.42 (0.17, 0.69)	0.25 (0.06, 0.52)	20	3
Adaptive validation (single)	0.57 (0.24, 0.90)	0.29 (0.02, 0.59)	10	2
First 200	0.50 (0.40, 0.61)	0.19 (0.11, 0.28)	200	66

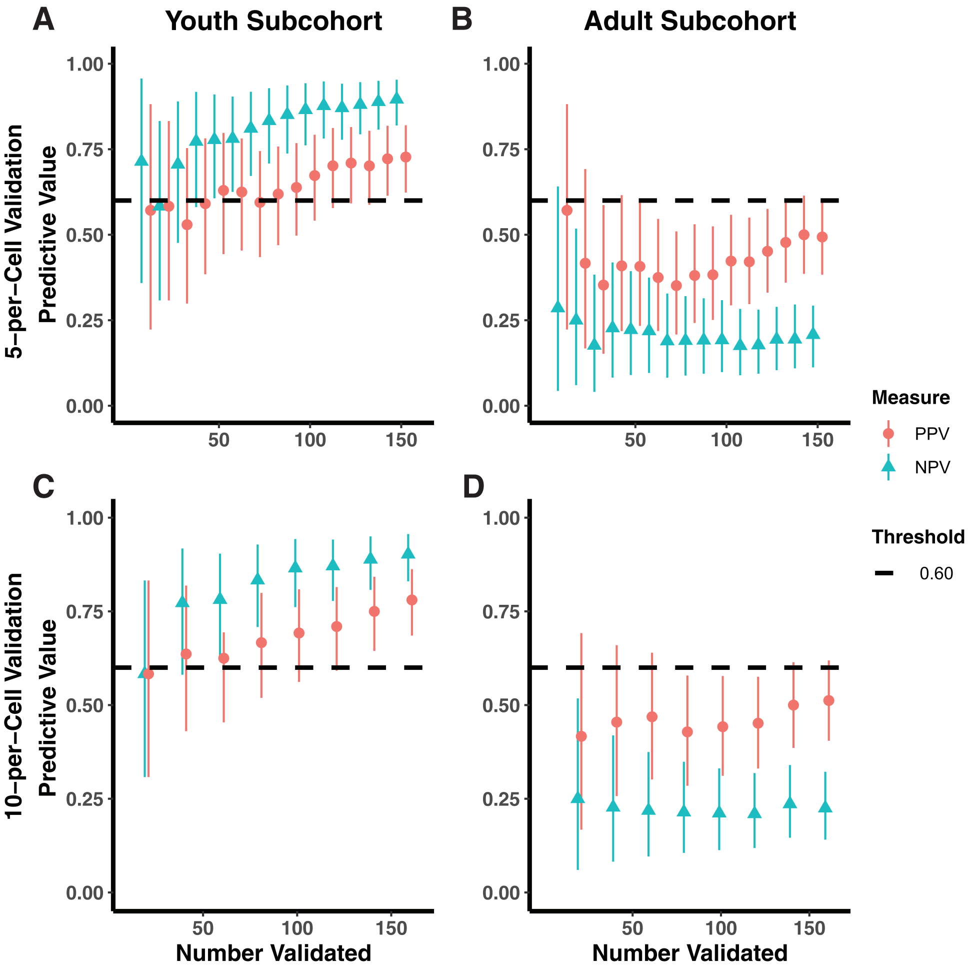

FIGURE 1. Adaptive validation using 5-per-cell validation scheme among STRONG (A) youth and (B) adult subcohorts, with comparison to the 10-per-cell validation scheme among the (C) youth and (D) adult subcohorts.

3 days into study enrollment and after 27 individuals were enrolled, with estimates for the PPV of 0.42 (95% CI = 0.17, 0.69) and NPV of 0.25 (95% CI = 0.06, 0.52). Similar results were obtained when we applied the 10-per-cell validation approach at each time interval. These results provided an early indication, although imprecise, that gold-standard measurement would need to be comprehensive to accurately classify the exposure of interest in this subcohort. In other words, the gold-standard measure was required for each cohort member because the mismeasured gender in the electronic medical record lacked validity.

Comparison to Validating the Full Cohort

We observed some differences in estimates of the PPV and NPV calculated from the adaptive validation approaches, compared with the overall estimates from the complete youth and adult subcohorts (Tables 1 and 2). In the youth cohort, PPV from the adaptive validation tended to underestimate the PPV obtained from the whole sample, whereas NPVs were roughly similar. In the adult cohort, PPVs in the adaptive designs were more similar to the PPV from the whole sample, whereas NPVs were more extreme. Because the validation study was stopped early in the adult cohort, these estimates

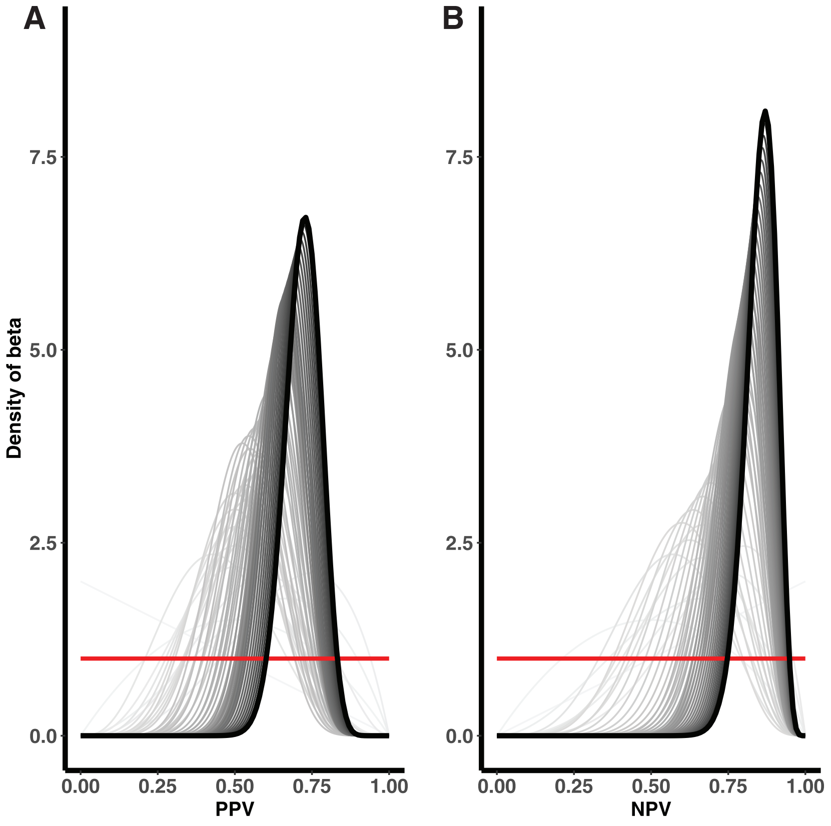

FIGURE 2. Single-person validation among STRONG youth cohort until the (A) PPV and (B) NPV were considered to have reached the predefined threshold for optimization.

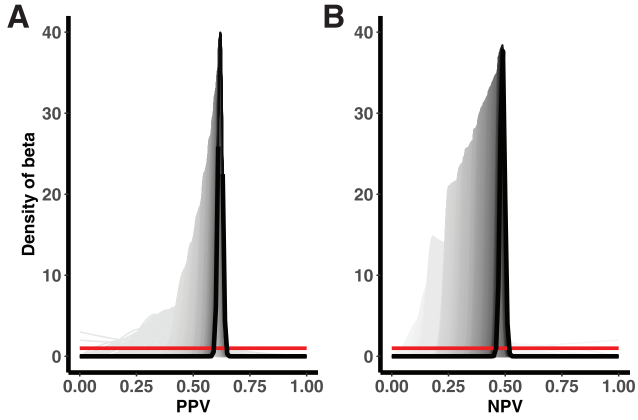

FIGURE 3. Single-person validation among complete STRONG adult cohort for the (A) PPV and (B) NPV.

were generally very imprecise. For both the youth and the adult subcohort, the differences in the results obtained from the adaptive validation approaches compared with the full cohort were indicative of a strong time-trend in the classification parameters within the STRONG parent cohort (Figures 2 and 3; eFigure 1; http://links.lww.com/EDE/B672).

Comparison With Other Approaches

We compared our method with another possible approach to validation with ongoing data collection. In the comparison scenario, we validated the first 200 enrolled members of the youth and adult subcohorts. In the youth subcohort, this method provided reasonable estimates, with an estimated PPV of 0.80 (95% CI = 0.71, 0.87) and NPV of 0.92 (95% CI = 0.84, 0.96) (Table 1). In contrast, in the adult subcohort, this method provided poor estimates (PPV = 0.50; 95% CI = 0.40, 0.61 and NPV = 0.19; 95% CI = 0.11, 0.28), which appeared to be due, in part, to the strong time-trend in classification parameters as noted earlier (Table 2). Both estimates were similar to those obtained from the adaptive validation approach, although they would have required additional expenditure of time and study resources.

DISCUSSION

We have developed an approach to validation study design suited to scenarios in which validation data are collected in real time and applicable to any parent epidemiologic study design. This method provides a valuable tool for prospective validation, allowing researchers to optimize the utility of fixed study resources when implementing validation studies. We demonstrated the ability of the method to determine when a validation study has reached the point that further validation efforts may be futile, based on a predefined threshold value. Using this design, the PPV and NPV calculated with the proposed study designs were generally comparable to those calculated from full validation efforts, albeit with less precision. However, fewer persons and less time were required for the adaptive validation studies.

Using this adaptive validation design can therefore save substantial costs and time, which can then be allocated to other research objectives. Iterative updating of classification parameters informs values of the parameters of interest and can be used as a marker for when sufficient information has been collected, indicating when validation efforts can stop. This process allows researchers to preferentially decide, in real time, how to appropriate resources to validation studies, with the ability to adapt based on the specific needs of their study. There is often a trade-off with internal validation studies—the possible sacrifice of sample size or other data collection of the parent study to obtain assurance that the variables of interest have adequate validity. Our method may enhance the ability of researchers to save on the resources allocated to validation, by thoughtfully validating over the study time period with periodic assessments of the performance of the variables mismeasured in the study. In this approach, the sample size necessary to achieve the specified threshold is potentially smaller than that required by other approaches, and validation can be completed before the end of follow-up of longitudinal studies with ongoing data accrual.

This proposed validation study design is amenable to alternative stopping rules based on prior knowledge, other criteria, or other validation study goals. For example, investigators may wish to rely exclusively on a threshold stopping rule, perhaps to identify futility. Alternatively, researchers may prefer a precision-based stopping rule, such as a prespecified width of the interval around the predictive value estimates that must be achieved before stopping, which can be easily incorporated into this method. A study may have inferential interest in the bias-adjusted estimate of association or occurrence. In this case, the uncertainty in the bias parameters would be incorporated in the bias-adjusted estimate and its uncertainty interval.[13,14] A stopping rule could be developed that ends the collection of both study data and validation data when the interval estimate around the main effect falls below a prespecified width. For all stopping rules, special consideration should be given to random error, which could inadvertently cause the validation study to stop too soon. Investigators may be interested in having stopping rules be evaluated only after the 50th individual has been enrolled, for instance. However, in our approach, we initiated validation efforts from the beginning of the study, which was particularly useful in the STRONG adult subcohort, where the method quickly identified that validation would be necessary for the entire study population. In this setting, the validation effort is really replacing the mismeasured variable (gender) with the gold standard (sex assigned at birth) to allow accurate classification of participants as transfeminine or transmasculine. Validation is, therefore, a misnomer; the validation substudy indicated that complete data would be required for the gold standard, and this is a new data collection effort.

The sampling approach for validation in this method can be readily applied to other studies and can be modified to suit the demands of individual applications. In this example, we relied on medical record abstraction for validation. Based on the number of persons who have their exposure status validated at each discrete time period, we included blocks of 5 and 10 as the approximate number of medical records that could be reviewed in 1 day by study staff. This example is applicable to other scenarios but can easily be extended to validation efforts that require other resources, such as additional procedures or tests. Consideration of the block size may also influence when stopping criteria will be met. For example, smaller block sizes allow for closer monitoring of the classification parameters, with the potential benefit of stopping earlier. As a result, careful consideration should be given to ensure that the validation study is not stopped too early and that the specified block size matches the practicalities of collecting the validation data (e.g., set the block size equal to the number of medical records that an abstractor could review in a typical day).

The current approach can be used to calculate estimates for the positive and negative predictive values, which are specific to the study from which they arise, but are not readily applied outside of the study population in which they are measured. Sensitivity and specificity are more easily transportable to other populations and also, therefore, more amenable to starting the adaptive validation analyses with an informative prior. The current approach does not allow for estimation of sensitivity and specificity as it conditions on the observed exposure status, but could be adapted to calculate these classification parameters. PPV and NPV are dependent on the prevalence of the measure of interest, so researchers should consider calculating PPV and NPV (of exposure) within strata of the outcome. However, this is difficult to accomplish in a study that prospectively validates data, as the outcome status may not be known at the time of validation. In validation studies that are conducted after the parent study is completed, this could be easily accomplished, and researchers may also want to sample from similar enrolment time periods, such as weekly or monthly intervals. For studies that evaluate the association between an exposure and multiple outcomes, additional considerations will apply, such as validation of exposure within strata of each outcome. Finally, the choice of stopping criteria can impact the estimates of the classification parameters and should be established carefully. The choice of a meaningful threshold is a subjective decision that should be informed before initiation of the validation study by substantive considerations. In the current study, we chose a threshold of 0.60 because PPVs or NPVs lower than this were deemed unacceptably low. A low threshold may, in some instances, result in imprecise classification parameters and researchers may want to specify a higher threshold or incorporate a precision criterion. Because of this, threshold criteria may be most useful when there is concern that a validation strategy may be unsuccessful. For demonstration of the method, we kept a general approach to highlight the situations in which validation could be halted to improve overall study accuracy and efficiency. Further work, involving simulation studies, should explore the magnitude of the bias from stopping validation too early.

Even though validation met the stopping threshold in the youth cohort, the estimates of the classification parameters differed slightly from the full cohort. This difference appeared to be related to a time-trend in the classification parameters, which was evident among both the youth and adult subcohorts (eFigure 1; http://links.lww.com/EDE/B672, Figure 3). The change in classification parameters over time may be due, in part, to the sociopolitical context surrounding transgender health, but it is conceivable that it could apply to other scenarios and measures that require validation. Note, however, that the time trend provides potentially useful information that might easily be missed if all validation data were collected and analyzed after the primary data collection was complete. Detection of a time trend may serve as its own stopping criterion, as the researchers would want to make sampling adjustments to conduct the adaptive validation approach within time periods that would more accurately capture the time trend in classification parameters. Furthermore, the corresponding bias adjustment would need to take the time trend into account, strengthening the validity of the bias-adjusted results. Other approaches do not allow for detection of a time trend, as illustrated by the fact that the trend was not, in fact, previously noted in the STRONG cohort analyses.

CONCLUSIONS

Our proposed adaptive validation design may be useful to calculate classification parameters as validation data accrue in epidemiologic studies, which can lead to effective and efficient conduct of validation substudies. Extending this proposed method to studies with multiple outcomes, measures of sensitivity and specificity, and using the bias-adjusted estimate of association to inform stopping rules will be important considerations for future development.

REFERENCES

1. Collins FS, Tabak LA. Policy: NIH plans to enhance reproducibility. *Nature*. 2014;505:612–613.
2. Lash TL, Olshan AF. Epidemiology announces the "Validation Study" submission category. *Epidemiology*. 2016;27:613–614.
3. Ehrenstein V, Petersen I, Smeeth L, et al. Helping everyone do better: a call for validation studies of routinely recorded health data. *Clin Epidemiol*. 2016;8:49–51.
4. Holcroft CA, Spiegelman D. Design of validation studies for estimating the odds ratio of exposure-disease relationships when exposure is misclassified. *Biometrics*. 1999;55:1193–1201.
5. Spiegelman D, Rosner B, Logan R. Estimation and inference for logistic regression with covariate misclassification and measurement error in main study/validation study designs. *J Am Stat Assoc.* 2000;95:51–61.
6. Spiegelman D, Gray R. Cost-efficient study designs for binary response data with Gaussian covariate measurement error. *Biometrics*. 1991;47:851–869.
7. Holford TR, Stack C. Study design for epidemiologic studies with measurement error. *Stat Methods Med Res*. 1995;4:339–358.
8. Greenland S. Variance estimation for epidemiologic effect estimates under misclassification. *Stat Med*. 1988;7:745–757.
9. Fayers PM, Ashby D, Parmar MK. Tutorial in biostatistics Bayesian data monitoring in clinical trials. *Stat Med*. 1997;16:1413–1430.
10. Quinn VP, Nash R, Hunkeler E, et al. Cohort profile: Study of Transition, Outcomes and Gender (STRONG) to assess health status of transgender people. *BMJ Open*. 2017;7:e018121.
11. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. *Bayesian Data Analysis*. Boca Raton, FL: Chapman and Hall/CRC; 2013.
12. Cronin-Fenton DP, Kjærsgaard A, Ahern TP, et al. Validity of Danish Breast Cancer Group (DBCG) registry data used in the predictors of breast cancer recurrence (ProBeCaRe) premenopausal breast cancer cohort study. *Acta Oncol*. 2017;56:1155–1160.
13. Lash TL, Fox MP, Fink AK. *Applying Quantitative Bias Analysis to Epidemiologic Data*. Boca Raton, FL: Springer Science & Business Media; 2011.
14. Rothman KJ, Greenland S, Lash TL. Bias analysis. In: *Modern Epidemiology*. Vol. 3. Philadelphia, PA: Wolters Kluwer Health/ Lippincott Williams & Wilkins; 2008.