Classification-specific predictive performance: A unified estimation and inference framework for multi-category tests

A. Gregory DiRienzo*, Elie Massaad & Hutan Ashrafian

Harbinger Health, Cambridge, MA, USA

* Corresponding author email: gdirienzo@harbinger-health.com

Abstract

Multi-Cancer testing with localization aims to detect signals from any of a set of targeted cancer types and predict the cancer signal origin from a single biological sample. Such tests have the potential to aid clinical decisions and significantly improve health outcomes. When used for multi-cancer screening in an asymptomatic population, these tests are referred to as multi-cancer early detection (MCED) tests. MCED testing has not yet achieved regulatory approval, reimbursement or broad clinical adoption. Some major reasons for this are that the clinical benefits and harms are not well understood, including the risk of unnecessary work-ups and false-reassurance from a negative test that could reduce uptake of standard-of-care screening. Part of this uncertainty stems from the use of clinically obtuse metrics to assess the test's clinical validity. Traditionally, performance of MCED tests has been quantified using aggregate measures, disregarding the joint distribution of cancer type, stage (both at intended-use incidence rates) and predicted cancer signal origin, thereby obscuring biological variability and underlying differences in the test's behavior and limiting insight into true effectiveness. Clinically informative evaluation of an MCED test's performance requires metrics that are specific to cancer type, stage and predicted cancer origin at expected incidence rates in the intended use population. In the context of a case-control sampling design, this paper derives analytical methods that allow for unbiased estimation of cancer-specific intrinsic accuracy, predicted cancer signal origin-specific predictive value and the marginal test classification distribution, each with corresponding valid confidence interval formulae. A simulation study is presented that evaluates performance of the proposed methodology and provides guidance for implementation. An application to a published MCED test dataset is given. The derived statistical analysis framework in general allows for estimation and inference for pointed metrics of a multi-category test that enables precisely informed decision-making, supports optimized trial designs across classical, digital, AI-driven, and hybrid stratified diagnostic

screening platforms, and facilitates informed healthcare decisions by clinicians, policymakers, regulators, scientists, and patients.

Key Words: CSO prediction, CSPP methodology, MCED tests; multi-category tests; predictive value; intrinsic accuracy; compound sum; confidence intervals

1. Introduction

Multi-category testing is used in many biomedical applications. Some examples include the Multi-category pharmacogenetic (PGx) test (1) that predicts a patient's response to various medications, the Multi-gene panel testing for cancer risk (2) that evaluates a patient's hereditary risk for multiple cancer types and Multi-cancer early detection (MCED) tests that use a single blood sample to ascertain the presence of any of a set of targeted cancer types (3). In this paper, we establish a general statistical framework that permits unbiased estimation and valid inference procedures for classification-specific predictive performance (CSPP) of multi-category tests. This methodology is shown to result in a more relevantly clear evaluation of a test's performance when establishing its validity. Because the potential impact of MCED tests on population health can be transformative, we center on this context to frame the development of our proposed CSPP statistical methodology. Furthermore, we focus on the class of MCED tests that predict the cancer signal origin (CSO) for a positive readout. These MCED tests inherently accommodate diverse populations of individuals composed of multiple clinical disease categories and attempt to classify subjects into one of multiple cancer types or a healthy state, resulting in a multi-dimensional confusion matrix rather than a conventional two-by-two table. This larger dimensionality requires a more complex evaluation of test performance compared to single-cancer binary tests. Conventional binary tests typically consider a two-way classification of disease versus no disease that leads to straightforward analyses of test performance (sensitivity, specificity, positive and negative predictive value). Deriving an appropriate analysis framework for MCED tests presents challenges for both defining clinically meaningful and actionable metrics and deriving corresponding unbiased estimation and valid inference procedures.

Performance metrics used to evaluate conventional binary tests that focus on the presence or absence of one disease allow a straight-forward clinically-focused interpretation. On the contrary, the broader applicability of MCED tests and their simultaneous assessment and categorization of multiple disease states does not afford a clear and obvious formulation of clinically meaningful and actionable performance metrics that can be used as a basis to assess clinical validity. Furthermore, the development of statistical estimation and inference techniques for such metrics requires

careful consideration of the underlying nuances in the observed data that are revealed upon this dimensionality expansion. Specifically, the increased dimensionality corresponding to MCED tests complicates:

i. the identification and definition of clinically meaningful and actionable performance metrics, that requires accommodating the potential for variation in performance across cancer types and stages, differences in cancer incidence rates, and dynamic clinical impacts;
ii. statistical estimation of and inference procedures for clinically meaningful and actionable metrics, resulting from increased complexity in appropriately quantifying uncertainty, the need to accommodate sparse data and to account for covariance among empirical cancer-specific performance measures;
iii. clinical interpretation and benefit/harm decision-making, because translating these complex multidimensional analytical frameworks into interpretable and actionable clinical insights for practitioners, policymakers, and regulatory agencies becomes significantly more challenging.

The purpose of this paper is to present a unified framework that identifies and defines clinically informative performance metrics for MCED tests along with analytical formulas that allow for unbiased estimation and valid confidence interval construction in the context of a case-control study design. Evidence generation from case-control studies is a necessary component of evaluating the performance of screening and diagnostic tests. At least two important reasons are:

i. compared to prospective cohort studies that enroll participants with unknown clinical outcome status, the procurement of case samples is structurally independent from the cohort of healthy controls that are enrolled, circumventing the requirement to recruit an overwhelmingly large number of individuals achieve a target number of cases based on an expected disease incidence and potentially lengthy clinical follow-up period.
ii. unlike prospective cohort studies that only perform follow-up workup to establish clinical outcome for those subjects that are a test positive at enrollment, case-control studies allow direct estimation of and inference for specificity and intrinsic accuracy metrics.

However, because cases are oversampled by design, case-control designs do not provide complete information on disease incidence, thereby precluding direct estimation of predictive value-type metrics and the marginal test readout distribution. We derive a framework that can be used to construct unbiased estimation and valid inference procedures for such metrics in this setting by using two approaches for providing information on incidence. When the intended use population is represented by that of a

national registry or published large population study, incidence rates from such databases can be employed. Alternatively, in settings where incidence information is not available from external sources, a sample-derived approach for obtaining conditional disease-type incidence that is paired with a posited overall value for disease incidence is proposed; this overall value can be varied in a sensitivity analysis if warranted.

After completion of one or more case-control studies, viable screening and diagnostic tools will generally have performance quantified using long-term and expensive prospective cohort studies conducted in the intended use population. The information obtained from case-control studies can have great value not only for aiding the design of future prospective studies but also assessing potential impact in clinical practice. For example, the result of the MCED test is generally intended to provide the practicing physician additional information to use as a part of the clinical decision-making process for first-line procedures. In particular, quantifying the MCED test's ability to correctly indicate the appropriate follow-up action on an individual and the test's corresponding contribution to the ultimate goal of improving patient outcomes in the intended use population are critical components (among others) for establishing the test's clinical utility (4, 5). This pointed information regarding the test's clinical validity is extremely valuable for patients, providers, regulators and payers before large and long-term prospective cohort studies can be completed and a more complete benefit and harm profile established (6, 7, 8).

Authors have advocated for the performance of MCED tests to be quantified separately for each cancer type under consideration in datasets representative of the intended use population (9). LeeVan and Pinsky (5) provide a review of the published clinical validity performance of several MCED tests as quantified by *conventionally* defined metrics. These conventional metrics include specificity, *crude* AUC and *crude* sensitivity (by cancer stage, by cancer type, but not both), and "TOO accuracy", defined as the proportion of correct CSO calls among test positive cases. Both AUC and sensitivity are described as "crude" because the predicted CSO is ignored in their definitions; only the "not test negative" event is used, thereby ignoring incorrect CSO predictions. Furthermore, because test performance generally depends on both cancer type and stage, conditioning on their joint distribution is necessary to uncover this heterogeneity. The metric "TOO accuracy" is devoid of a clear population analog, rendering a vacuous interpretation because the conditioning event is both a cancer case from a defined targeted panel and a not negative test result; furthermore, intended-use population cancer type and stage incidence is ignored in its calculation. Alternatively, we propose metrics for establishing clinical validity that quantify the *intrinsic accuracy* of each cancer type (by stage) evaluated by the MCED test in addition to the *predictive value positive* for each CSO readout category. Obtaining intrinsic accuracy metrics for each

cancer type (by stage) quantifies an MCED test's *hypothetical* ability to control both false negative and incorrect CSO classification results; "hypothetical" here describes metrics that condition on an unobservable event (e.g. cancer status at test administration). Predictive value metrics for each CSO readout category address the corresponding false positive error and associated risk of incorrectly informing decisions on ensuing clinical follow-up indicated by the test result. Predictive value metrics are *actionable* because they condition only on an observable event (test readout).

Several authors have noted that the clinical validity of an MCED test that is quantified using a case-control study does not necessarily reflect the test's performance in an asymptomatic screening population (9, 10, 11). Some reasons for this include the potential for *spectrum* bias in the composition of cases and controls (9), as well as a presumed shift to larger early-stage cancer incidence in an asymptomatic screening population (10, 11). Moreover, establishing clinical validity even with a prospective study in an asymptomatic screening population is difficult because of the challenge in determining the absence/presence of preclinical cancer and cancer stage for all study participants at test administration (12). Pinsky, Lange and Etzioni (12) present a natural history modeling approach as a basis to derive stage-specific sensitivity estimators in prospective (and retrospective) studies that can reduce bias for true sensitivity targets compared to conventional "empirical sensitivity" estimators that use the "detection method". .

Section 2 briefly reviews standard metrics for 2 x 2 tables and Section 3 expands these definitions to derive clinically meaningful and actionable metrics corresponding to a *(J+1)* x *(K+1)* joint distribution of disease state versus MCED test readout. In Section 4, a proposed framework that allows the potential for unbiased estimation and valid inference in general settings for various types of metrics for use in pointed quantification of the test's clinical validity are provided in the context of a case-control sampling design. First, estimators of intrinsic accuracy for a given cancer state are derived that recognize these quantities are in fact compound random variables because, although the total number of cases is predetermined, generally, the realized number of cases of a specific cancer type (and stage) varies randomly. In addition, metrics are defined for the false-negative probability for each cancer state as well as an overall intrinsic accuracy metric. Estimation and inference procedures are also provided for predictive value negative and, separately for each CSO readout category, predictive value positive metrics; a method for obtaining overall predictive value positive is also proposed. Estimation and inference for the marginal test readout category distribution is also presented; this can be useful for evaluating the impact of the MCED test on outcomes in the intended use population and for aiding the design of future prospective studies, for example by informing the number of subjects needed to be enrolled to

realize an expected number of readouts in each CSO category. An approach is proposed to adjust false-positive counts for tests with very high specificity to stabilize estimation and inference for predictive value metrics and the marginal test distribution in the presence of sparse data. Inference for each type of metric is performed using the delta-method on the logit transform to obtain valid confidence intervals for the corresponding population parameter in the unit interval. Stratified analyses are discussed in Section 5 and Section 6 presents a numerical simulation study that evaluates performance of the proposed methodology in various clinical use settings. Analysis of a real-world case-control study is provided in Section 7 and Section 8 concludes with a discussion, where several topics are addressed, including the use of case-control studies for assessing clinical validity in an asymptomatic screening population.

2. Standard metrics for binary tests and disease states using case-control sampling

With a binary test readout "+" or "-" and binary disease outcome (Disease ("Case") or no Disease ("Control")), the two-way joint distribution of test and outcome is:

Table 1:

A case-control design fixes in advance each the number of cases and controls that are included in the study; under a random sampling model for each group, a binomial model applies for the number of cases that test positive and the number of controls that test positive. That is, $n_{01} \sim Binomial\{N_0, P(Test + | \, no \; Disease)\}$ and $n_{11} \sim Binomial\{N_1, P(Test + | \, Disease)\}$, where $P(X)$ is the probability measure of the random variable X arising from the target intended use population, denoted by $\boldsymbol{\mathcal{P}}$.

The test performance metrics sensitivity and specificity are defined as:

$$\textit{sensitivity (SE)} = P(Test + | \, Disease)$$

$$\textit{specificity (SP)} = P(Test - | \, no \; Disease) = 1 - \; P(Test + | \, no \; Disease)$$

and are directly estimable from this sampling model. The corresponding maximum likelihood estimators are the binomial proportions $\widehat{SE} = \; n_{11}/n_{1+}$ and $\widehat{SP} = \; n_{00}/n_{0+}$, respectively.

The positive (PPV) and negative (NPV) predictive value of the test are not directly estimable with the case-control sampling model since cases are oversampled from the intended use population $\boldsymbol{\mathcal{P}}$. To obtain estimates of these metrics, *Bayes rule* and the *Law of total probability* need to be used along with a working value for the disease

incidence, denoted $P(D)$. The disease incidence is the expected number of cases per 1 person-year of follow-up in individuals from $\boldsymbol{\mathcal{P}}$, with $1 - P(D) = P(\overline{D})$ since $D \cup \overline{D} = \boldsymbol{\mathcal{P}}$ and the events of Disease (D) and no Disease ($\overline{D}$) are mutually exclusive, that is, $D \cap \overline{D} = \{\emptyset\}$. Using Bayes' rule, PPV = $P(D \mid Test+) = P(Test+ \mid D)P(D)/P(Test+) = SE(P(D))/P(Test+)$. Using the Law of total probability, $P(Test+) = SE\big(P(D)\big) + (1 - SP)(1 - P(D))$. Similarly, NPV = $P(\overline{D} \mid Test-) = SP(1 - P(D))/\{SP(1 - P(D)) + (1 - SE)P(D)\}$. Methods exist for confidence interval construction for PPV and NPV in this 2 x 2 setting shown in Table 1 (13).

3. Metrics for multi-category tests and disease states

Suppose that there are $J > 2$ mutually exclusive disease states, labeled $D_j, j = 1, \ldots, J$; let D_0 denote the no disease state. Then $\bigcup_{j=0}^{J} D_j = \boldsymbol{\mathcal{P}}$, with $D_j \cap D_\ell = \{\emptyset\}, (j \neq \ell)$, and $\sum_{j=0}^{J} P\big(D_j\big) = 1$. Each disease state may include one or more distinct disease types and/or stages that share a common implication regarding the test's clinical validity; for example, a group of disease types that share a common first-line clinical workup. Although it is possible in rare occurrences that an individual from $\boldsymbol{\mathcal{P}}$ has simultaneous membership to two distinct disease states, we assume this event has probability 0, although the methodology presented can accommodate this situation if necessary by defining this union as a separate disease state with non-zero probability.

In addition to a Negative category, the test has a readout category corresponding to *K* $(K \leq J)$different disease states (CSOs). Each of the test readout categories are denoted $Test = k : k = 0, \ldots, K$, with $Test = 0$ denoting the Negative readout and test positive categories 1 to *K* assumed to be a one-to-one correspondence to the disease states D_1 to D_K. This defined setting assumes one and only one CSO prediction is provided for a positive test readout; alternative test configurations are presented in the Discussion. There may be disease states, labeled D_{K+1} to D_J, that do not have a corresponding CSO test readout category; further details of this setting are provided in the Discussion. It is assumed that the data arise as a random sample of N_0subjects with no disease and $(N - N_0)$ cases from intended use population $\boldsymbol{\mathcal{P}}$. Table 2 below shows a depiction of the joint distribution in this setting.

Table 2:

Two general types of metrics are considered for conveying useful information regarding the relationship between test and disease. *Intrinsic accuracy* measures the conditional probability of the *target* test result among individuals with a given disease state;

Predictive value measures the conditional probability of the *target* disease state among individuals with a given test readout category.

Intrinsic accuracy corresponding to disease state $j = 0, \dots, K$ is defined as:

$$A_j = \frac{P(Test = j, D_j)}{\sum_{k=0}^{K} P(Test = k, D_j)} = \frac{P(Test = j, D_j)}{P(D_j)}, j = 0, \dots, K.$$

The intrinsic accuracy metric can be written as $A_j = P\big(\, Test = j \mid D_j \,\big)$. With a binary disease and test, A_0 denotes specificity and A_1 denotes sensitivity.

Predictive value *positive* corresponding to test readout category $k = 1, \dots, K$ is:

$$PVP_k = \frac{P(Test = k, D_k)}{\sum_{j=0}^{J} P(Test = k, D_j)} = \frac{P(Test = k, D_k)}{P(Test = k)}, k = 1, \dots, K.$$

Predictive value *negative* corresponding to test readout category $k = 0, \dots, K$ is:

$$PVN_k = \frac{P(Test = k, D_0)}{\sum_{j=0}^{J} P(Test = k, D_j)} = \frac{P(Test = k, D_0)}{P(Test = k)}, k = 0, \dots, K.$$

The predictive value metrics can be written as $PVP_k = P(D_k \mid Test = k)$ and $PVN_k = P(D_0 \mid Test = k)$. With a binary disease and test, PVP_1 is PPV and PVN_0 is NPV.

To facilitate contextualization of the methodology developed in this paper, presented in the table directly below is data published for a case-control study used to demonstrate the performance of a blood-based MCED test (14). The generation of this table and detailed analyses are described in Section 7 below. Some important observations are:

i. although the total number of cancer cases are considered fixed by design, the number of cases of each type and stage is random;
ii. for a given cancer type, three types of classifications are possible: Negative, positive with incorrect CSO classification, positive with correct CSO classification;
iii. clinically actionable information necessitates establishing predictive value positive separately for each CSO readout and needs to address the oversampling of cases and the sparse false positive counts in the Control group.

It is clear that, in order to properly interpret such large-dimensional confusion matrices, it is required to formulate focused performance metrics that target key pieces of information that can be of value in aiding clinical decision making.

Figure 6 B. from Liu et al. (2020) (with modification):

4. Estimation and inference for performance metrics for multi-category tests and disease states using case-control sampling

For a case-control design, both the number of controls, N_0, and the number of cases, $N - N_0 = N_1$ are fixed in advance by the design. However, the number of cases of a particular disease state, $n_{j+}, j = 1, \ldots, J$, are not in general fixed in advance in the study design; instead, these arise as random variables that are dependent on the underlying disease incidence and study sampling process. In this setting, under random sampling of N_1 cases from $\boldsymbol{\mathcal{P}}$, the $\{n_{j+}: j = 1, \ldots, J-1\}$ constitute a multinomial random vector with parameters N_1 and $\{p_j: j = 1, \ldots, J-1\}$, with $p_j = P(D_j \mid D)$ and $p_J = 1 - \sum_{j=1}^{J-1} P(D_j \mid D)$. Maximum likelihood estimates for $P(D_j \mid D)$ are $\hat{p}_j = n_{j+}/N_1$ and $\hat{p}_J = 1 - \sum_{j=1}^{J-1} \hat{p}_j$.

Estimation and inference for predictive value metrics require values for disease incidence $P(D_j)$. Two approaches are considered for this purpose. If a national registry database or published large population study exists for the intended use population, values of $P(D_j)$ can be sourced from it. If the observed data in Table 2 is truly obtained via a conditional random sample of N_0 controls and N_1 cases from $\boldsymbol{\mathcal{P}}$, because $P(D_j) = P(D_j, D) = P(D_j \mid D)P(D)$, an estimate of $P(D_j)$ may be obtained from multiplying an estimate of $P(D_j \mid D)$ obtained from the case-control dataset and a proposed value for $P(D)$, possibly obtained from an external source or subject matter experts if none exists. Both of these approaches to obtain $P(D_j)$ will be evaluated below.

Define the shorthand notation for test readout category $T_k \equiv (Test = k)$. When $j = 0$, the maximum likelihood estimator for $A_0 = P(T_0 \mid D_0)$ is the Binomial proportion n_{00}/N_0. Similarly, the false-positive error corresponding to each test category is defined as $\beta_k = P(T_k \mid D_0), k = 1, \ldots, K$, with corresponding maximum-likelihood estimators $\hat{\beta}_k = n_{0k}/N_0$. Standard confidence intervals for A_0 and β_k can be obtained, for example the Mid-P approach (15); this approach is used in the Simulation study and Data analysis sections below. In settings with very high specificity, e.g. screening tests, the counts n_{0k} can be sparse when the number of positive test categories *K* is large relative to N_0, this can make direct estimation of $P(T_k \mid D_0), k = 1, \ldots, K$, challenging. Approaches to address this challenge are proposed in section 4.4 below.

4.1 Estimation and inference for intrinsic accuracy metrics

To derive an unbiased estimator for $A_{jk} = P(T_k \mid D_j), j = 1, \ldots, J, k = 0, \ldots, K$, define the Bernoulli random variable $Y_i^{(jk)} = I(subject\ i\ from\ group\ D_j\ has\ Test = k)$, where $I(a)$ is the indicator function, assuming the value 1 for the event a true, 0 otherwise. Note

$Y_i^{(jk)}$has expectation $P(T_k \mid D_j)$. The sum $n_{jk} = \sum_{i=1}^{n_{j+}} Y_i^{(jk)}$is a *compound random variable* since both n_{j+} and $Y_i^{(jk)}$ are random variables and $Y_i^{(jk)}$is considered independent of n_{j+}. When the $\{n_{j+}: j = 1, \ldots, J-1\}$ arise from multinomial sampling, the random count n_{j+} has a Binomial distribution with parameters $N_1$and $P(D_j \mid D)$.

Define $\tilde{A}_{jk} = \left(\frac{1}{n_{j+}}\right) \sum_{i=1}^{n_{j+}} Y_i^{(jk)}$. Using the convention that $0/0 \equiv 0$, it can be shown that $E(\tilde{A}_{jk}) = E\left(Y_i^{(jk)}\right) P(n_{j+} > 0)$ and

$$V(\tilde{A}_{jk}) = E^2\left(Y_i^{(jk)}\right) P(n_{j+} > 0)\{1 - P(n_{j+} > 0)\} + V\left(Y_i^{(jk)}\right) E\left(\frac{1}{n_{j+}} \mid n_{j+} > 0\right) P(n_{j+} > 0),$$

where $E(X)$ and $V(X)$ denote expectation and variance of the random variable X with respect to $P(X)$. Consider estimating A_{jk} by $\hat{A}_{jk} = \tilde{A}_{jk}/P(n_{j+} > 0)$. The expectation of $\hat{A}_{jk}$ is:

$$E\left(\frac{\tilde{A}_{jk}}{P(n_{j+} > 0)}\right) = \frac{E(\tilde{A}_{jk})}{P(n_{j+} > 0)} = E\left(Y_i^{(jk)}\right) = P(T_k \mid D_j)$$

and the variance of $\hat{A}_{jk}$ is

$$\begin{aligned}\sigma_{jk}^2 = V(\hat{A}_{jk}) &= V(\tilde{A}_{jk})/P^2(n_{j+} > 0) \\ &= \frac{1}{P(n_{j+} > 0)}\Bigg[P^2(T_k \mid D_j)\{1 - P(n_{j+} > 0)\} \\ &\quad + P(T_k \mid D_j)\{1 - P(T_k \mid D_j)\}E\left(\frac{1}{n_{j+}} \mid n_{j+} > 0\right)\Bigg].\end{aligned}$$

The sample analog for $V(\hat{A}_{jk})$, denoted by $\hat{\sigma}_{jk}^2$, is obtained by substituting $\hat{A}_{jk}$ for $P(T_k \mid D_j)$ in the expression directly above. The quantities $\hat{A}_{jk}$ and $\hat{\sigma}_{jk}^2$ rely on a known distribution for n_{j+}. In practice, $\hat{p}_j$ is substituted for $P(D_j \mid D)$ in the Binomial distribution for n_{j+}. Because N_1 is typically considered large, errors in this approximation are assumed negligible; this claim is supported in the simulation study.

Consider deriving a confidence interval for $\log[\text{odds}\{P(T_k \mid D_j)\}] = \text{logit}\{P(T_k \mid D_j)\}$, which for both analytical and numerical considerations is preferred over direct evaluation of $P(T_k \mid D_j)$. Write

$$\text{logit}\{P(T_k \mid D_j)\} = \log\left\{\frac{P(T_k \mid D_j)}{\sum_{\ell=0}^{K} I(\ell \neq k) P(T_\ell \mid D_j)}\right\}.$$

An estimator for the above quantity substitutes $\hat{A}_{j\ell}$ for $P(T_\ell \mid D_j)$, which equals

$$\mathrm{L}\left(\tilde{A}_{jk}\right) = \log\left\{\frac{\tilde{A}_{jk}}{\sum_{\ell=0}^{K} I(\ell \neq k)\tilde{A}_{j\ell}}\right\} = \log\left(\frac{n_{jk}/n_{j+}}{\sum_{\ell=0}^{K} I(\ell \neq k)\frac{n_{j\ell}}{n_{j+}}}\right) = \log\left(\frac{n_{jk}/n_{j+}}{1-\frac{n_{jk}}{n_{j+}}}\right)$$
$$= \log\left(\frac{\tilde{A}_{jk}}{1-\tilde{A}_{jk}}\right).$$

The approximate variance of $\mathrm{L}\left(\tilde{A}_{jk}\right)$ can be obtained using the delta-method, which is evaluated to equal:

$$V\left\{\mathrm{L}\left(\tilde{A}_{jk}\right)\right\}$$
$$= \frac{P\left(n_{j+}>0\right)\left[P\left(T_k \mid D_j\right)\left\{1-P\left(n_{j+}>0\right)\right\} + \left\{1-P\left(T_k \mid D_j\right)\right\}E\left(\frac{1}{n_{j+}} \mid n_{j+}>0\right)\right]}{P\left(T_k \mid D_j\right)\left\{1-P\left(T_k \mid D_j\right)\right\}^2}.$$

An estimator for $V\left\{\mathrm{L}\left(\tilde{A}_{jk}\right)\right\}$ is obtained by substituting $\hat{A}_{jk}$ for $P\left(T_k \mid D_j\right)$ and $\hat{p}_j$ for $P\left(D_j \mid D\right)$ in the Binomial distribution for n_{j+}, denoted $\hat{V}\left\{\mathrm{L}\left(\tilde{A}_{jk}\right)\right\}$.

A $100(1-\alpha)\%$ Wald-type valid confidence interval for $\mathrm{logit}\left\{P\left(T_k \mid D_j\right)\right\}$ is:

$$\mathrm{L}\left(\tilde{A}_{jk}\right) \pm z_{1-\alpha/2}\sqrt{\hat{V}\left\{\mathrm{L}\left(\tilde{A}_{jk}\right)\right\}}.$$

The anti-logit transform of this confidence interval yields the corresponding one for $A_{jk.}$

The intrinsic accuracy estimator of A_j corresponding to disease state $j = 1, \ldots, K$ is $\hat{A}_{jj} = \tilde{A}_{jj}/P\left(n_{j+}>0\right)$ and the estimated logit transform is $\mathrm{L}\left(\tilde{A}_{jj}\right)$. The metric $1-A_j$ for $j = 1, \ldots, K$, encompasses two types of misclassification error:

i. cases that obtain a Negative test readout, and
ii. cases that obtain a positive test readout but the incorrect cancer signal origin classification.

To enable separation of these two types of misclassification errors, define the metric $A_j^0 = P\left(T_0 \mid D_j\right)$; this is the false Negative classification error corresponding to disease state $j = 1, \ldots, J$. To obtain an estimator and corresponding $100(1-\alpha)\%$ confidence interval for A_j^0, the Bernoulli variable $Y_i^{(j0)}$ is used in the inference procedure described above, with ensuing notational changes resulting from setting the value $k = 0$. Note that the value $1-A_j^0 = 1-P(T_0 \mid D_j)$ is below referred to as "crude sensitivity" for disease state j; it is the probability of not a Negative test and encompasses errors of type ii. defined directly above.

For an overall intrinsic accuracy measurement, the aggregate proportion of subjects across disease states $1, \ldots, K$ that obtain the correct test classification is calculated as:

$$\sum_{k=1}^{K} \hat{A}_{kk} \left\{ \frac{P(D_k \mid D)}{\sum_{\ell=1}^{K} P(D_\ell \mid D)} \right\},$$

estimates for $P(D_k \mid D)$ can be obtained as described above. This aggregate metric does not transcend information for use at a per-subject level.

4.2 Estimation and inference for predictive value metrics

For analysis of predictive value metrics, write $P(T_k, D_j) = P(T_k \mid D_j)P(D_j)$ and $P(D_j) = P(D_j, D) = P(D_j \mid D)P(D), j = 1, \ldots, J$. Because $P(T_k \mid D_j)$ and $P(D_j \mid D)$ are directly estimable with a case-control design and random sampling of cases and controls from $\boldsymbol{\mathcal{P}}$, the general form for PVP_k is:

$$PVP_k = \frac{P(T_k \mid D_k)P(D_k \mid D)P(D)}{P(T_k \mid D_0)\{1 - P(D)\} + P(D)\sum_{j=1}^{J} P(T_k \mid D_j)P(D_j \mid D)}, \quad k = 1, \ldots, K.$$

Similarly,

$$PVN_k = \frac{P(T_k \mid D_0)\{1 - P(D)\}}{P(T_k \mid D_0)\{1 - P(D)\} + P(D)\sum_{j=1}^{J} P(T_k \mid D_j)P(D_j \mid D)}, \quad k = 0, \ldots, K.$$

A value for overall disease incidence $P(D)$ is assumed to be obtained from an external source, such as a population disease registry, published large population study or subject matter experts, and is considered fixed for inferential purposes. Unbiased estimators for PVP_k and PVN_k and their variances derived below may be obtained by substitution of $\hat{A}_{jk}$ for $P(T_k \mid D_j)$ and n_{0k}/N_0 for $P(T_k \mid D_0)$. Two settings will be considered for obtaining values for $P(D_j \mid D)$, either sourcing from a national registry or plugging in the estimator $\hat{p}_j$.

To conduct valid inference for the population predictive value, the logit transformation is used along with the delta-method approach for variance formulation. Define the logit transform of PVP_k by:

$$U(\boldsymbol{\varphi_k}) = \log\left\{ \frac{P(T_k \mid D_k)P(D_k \mid D)P(D)}{R_U(\boldsymbol{\varphi_k})} \right\}, k = 1, \ldots, K,$$

where

$R_U(\boldsymbol{\varphi_k}) = P(\,T_k \mid D_0\,)\{1 - P(D)\}$
$+ P(D)\sum_{j=1}^{J-1} I(j \neq k)\, P(T_k \mid D_j)P(D_j \mid D) + P(D)\, I(J \neq k)P(T_k \mid D_J)\{1 - \sum_{j=1}^{J-1} P(D_j \mid D)\}.$

The logit transformation of PVN_k , $k = 0, \ldots, K$, is:

$$W(\boldsymbol{\varphi_k}) = \log\left\{\frac{P(\,T_k \mid D_0\,)\{1 - P(D)\}}{P(D)\sum_{j=1}^{J-1} P(T_k \mid D_j)P(D_j \mid D) + P(D)P(T_k \mid D_J)\left\{1 - \sum_{j=1}^{J-1} P\big(\,D_j \mid D\,\big)\right\}}\right\}.$$

Here, $\boldsymbol{\varphi_k}$ is the $(2J)$-column vector of parameters:

$$\boldsymbol{\varphi_k} = \{\,P(T_k \mid D_0\,), P(\,T_k \mid D_1\,), \ldots, P(T_k \mid D_J), P(D_1 \mid D), \ldots, P(D_{J-1} \mid D)\}^t$$

with $\boldsymbol{a}^t$ denoting the transpose of the matrix $\boldsymbol{a}$. Denote the sample analog of $\boldsymbol{\varphi_k}$ by $\widehat{\boldsymbol{\varphi}}_{\boldsymbol{k}}$ that substitutes the corresponding estimators for components of $\boldsymbol{\varphi_k}$ as defined above.

Using the delta-method, the asymptotic variance of the sample version of $\mathrm{logit}(PVP_k)$, denoted $U(\widehat{\boldsymbol{\varphi}}_{\boldsymbol{k}})$, that substitutes $\hat{p}_j$ for $P\big(\,D_j \mid D\,\big)$ is:

$$V\{U(\widehat{\boldsymbol{\varphi}}_{\boldsymbol{k}})\} = \{U'(\boldsymbol{\varphi_k})\}^t\, V(\widehat{\boldsymbol{\varphi}}_{\boldsymbol{k}} - \boldsymbol{\varphi_k})\{U'(\boldsymbol{\varphi_k})\}\,,$$

where the column vector $U'(\boldsymbol{\varphi_k})$ is:

$$\{U'(\boldsymbol{\varphi_k})\}^t = \left(\frac{\partial U(\boldsymbol{\varphi_k})}{\partial P(T_k \mid D_0)}, \frac{\partial U(\boldsymbol{\varphi_k})}{\partial P(T_k \mid D_1)}, \ldots, \frac{\partial U(\boldsymbol{\varphi_k})}{\partial P(T_k \mid D_J)}, \frac{\partial U(\boldsymbol{\varphi_k})}{\partial P(D_1 \mid D)}, \ldots, \frac{\partial U(\boldsymbol{\varphi_k})}{\partial P(D_{J-1} \mid D)}\right)$$

and $V(\widehat{\boldsymbol{\varphi}}_{\boldsymbol{k}} - \boldsymbol{\varphi_k})$ is the $(2J \times 2J)$ variance-covariance matrix of $\widehat{\boldsymbol{\varphi}}_{\boldsymbol{k}}$. The components of $\{U'(\boldsymbol{\varphi_k})\}^t$ and $V(\widehat{\boldsymbol{\varphi}}_{\boldsymbol{k}} - \boldsymbol{\varphi_k})$ are provided in the Appendix.

When the terms $P\big(\,D_j \mid D\,\big)$ are considered fixed, for example when obtained from a population registry, the asymptotic variance of the corresponding version of $U(\widehat{\boldsymbol{\varphi}}_{\boldsymbol{k}})$ is

$$V_0\{U(\widehat{\boldsymbol{\varphi}}_{\boldsymbol{k}})\} = \{U_0'(\boldsymbol{\varphi_k})\}^t\, V_0(\widehat{\boldsymbol{\varphi}}_{\boldsymbol{k}} - \boldsymbol{\varphi_k})\{U_0'(\boldsymbol{\varphi_k})\}\,,$$

where U_0' and V_0 are as U' and V except with all components involving $P\big(\,D_j \mid D\,\big)$ set to equal 0.

Similarly, using the delta-method, the asymptotic variance of the sample version of $\mathrm{logit}(PVN_k)$, denoted $W(\widehat{\boldsymbol{\varphi}}_{\boldsymbol{k}})$, that substitutes $\hat{p}_j$ for $P\big(\,D_j \mid D\,\big)$ is:

$$V\{W(\widehat{\boldsymbol{\varphi}}_{\boldsymbol{k}})\} = \{W'(\boldsymbol{\varphi_k})\}^t\, V(\widehat{\boldsymbol{\varphi}}_{\boldsymbol{k}} - \boldsymbol{\varphi_k})\{W'(\boldsymbol{\varphi_k})\}\,,$$

where the column vector $W'(\boldsymbol{\varphi_k})$ is:

$$\{W'(\boldsymbol{\varphi_k})\}^t = \left(\frac{\partial W(\boldsymbol{\varphi_k})}{\partial P(T_k \mid D_0)}, \frac{\partial W(\boldsymbol{\varphi_k})}{\partial P(T_k \mid D_1)}, \ldots, \frac{\partial W(\boldsymbol{\varphi_k})}{\partial P(T_k \mid D_J)}, \frac{\partial W(\boldsymbol{\varphi_k})}{\partial P(D_1 \mid D)}, \ldots, \frac{\partial W(\boldsymbol{\varphi_k})}{\partial P(D_{J-1} \mid D)}\right).$$

The components of $\{W'(\boldsymbol{\varphi_k})\}^t$ are provided in the Appendix. When the $P(D_j \mid D)$ are fixed, the asymptotic variance of $W(\widehat{\boldsymbol{\varphi}}_{\boldsymbol{k}})$ takes the form

$$V_0\{W(\widehat{\boldsymbol{\varphi}}_{\boldsymbol{k}})\} = \{W_0'(\boldsymbol{\varphi_k})\}^t \, V_0(\widehat{\boldsymbol{\varphi}}_{\boldsymbol{k}} - \boldsymbol{\varphi_k})\{W_0'(\boldsymbol{\varphi_k})\},$$

Where W_0' is as W' except with all components involving $P(D_j \mid D)$ set to equal 0.

When estimating $P(D_j \mid D)$ with $\hat{p}_j$, a $100(1-\alpha)\%$ a valid Wald-type confidence interval for $\text{logit}\{PVP_k\}$ is

$$U(\widehat{\boldsymbol{\varphi}}_{\boldsymbol{k}}) \pm z_{1-\alpha/2}\sqrt{\hat{V}\{U(\widehat{\boldsymbol{\varphi}}_{\boldsymbol{k}})\}}, k = 1, \ldots, K,$$

where $\hat{V}\{U(\widehat{\boldsymbol{\varphi}}_{\boldsymbol{k}})\}$ arises from substituting sample analogs for population parameters in $V\{U(\widehat{\boldsymbol{\varphi}}_{\boldsymbol{k}})\}$; when $P(D_j \mid D)$ is considered fixed, $\hat{V}_0\{U(\widehat{\boldsymbol{\varphi}}_{\boldsymbol{k}})\}$ replaces $\hat{V}\{U(\widehat{\boldsymbol{\varphi}}_{\boldsymbol{k}})\}$. A $100(1-\alpha)\%$ Wald-type confidence interval for $\text{logit}\{PVN_k\}$ can analogously be constructed. The anti-logit transform of the corresponding confidence interval yields the one for PVP_k or PVN_k.

An *overall* Predictive value *positive* metric corresponding to all positive Test readout categories $k = 1, \ldots, K$ is defined as:

$$PVP^* = \frac{\sum_{k=1}^{K} P(T_k, D_k)}{\sum_{k=1}^{K} P(T_k)} = \sum_{k=1}^{K} P(D_k \mid T_k)\left\{\frac{P(T_k)}{\sum_{\ell=1}^{K} P(T_\ell)}\right\}$$
$$= \frac{\sum_{k=1}^{K} P(T_k \mid D_k)P(D_k \mid D)P(D)}{\sum_{k=1}^{K}\left[\sum_{j=1}^{J} P(T_k \mid D_j)P(D_j \mid D)P(D) + P(T_k \mid D_0)\{1 - P(D)\}\right]}.$$

As was the situation when considering an aggregate intrinsic accuracy metric, an aggregate predictive value metric has no clear population analog to facilitate its interpretation at a per-subject level, and its development is not further considered.

4.3 Estimation and inference for the marginal test category distribution

Establishing unbiased estimation and valid inference procedures for the marginal probabilities $P(T_k) = P(Test = k), k = 0, \ldots, K$, provides useful information for interpreting the test's potential impact in the population and for designing prospective cohort studies. Inference for $P(T_k)$ can be greatly simplified if the full distribution in Table 2 is collapsed as shown in Table 3.

Table 3:

Write $P(T_k) = P(T_k \mid D_0)\{1 - P(D)\} + P(T_k \mid D)P(D), k = 0, \ldots, K$. The logit transformation of $P(T_k)$ is:

$$Q(\boldsymbol{\vartheta}_k) = \log\left[\frac{P(T_k \mid D_0)\{1 - P(D)\} + P(T_k \mid D)P(D)}{\{1 - P(T_k \mid D_0)\}\{1 - P(D)\} + \{1 - P(T_k \mid D)\}P(D)}\right] = \log\left(\frac{B_k^0}{B_k^1}\right),$$

where $\boldsymbol{\vartheta}_k^t = \big(P(T_k \mid D_0), P(T_k \mid D)\big)$. The sample analog of $\boldsymbol{\vartheta}_k$ is denoted by $\widehat{\boldsymbol{\vartheta}}_k$, that substitutes the Binomial proportions n_{0k}/N_0 for $P(T_k \mid D_0)$ and $(n_{+k} - n_{0k})/N_1$ for $P(T_k \mid D)$. Using the delta-method, the asymptotic variance of $Q(\widehat{\boldsymbol{\vartheta}}_k)$, the sample version of logit$\{P(T_k)\}$, is $V\{Q(\widehat{\boldsymbol{\vartheta}}_k)\} = \{Q'(\boldsymbol{\vartheta}_k)\}^t\, V(\widehat{\boldsymbol{\vartheta}}_k - \boldsymbol{\vartheta}_k)\{Q'(\boldsymbol{\vartheta}_k)\}$,

where the column vector $Q'(\boldsymbol{\vartheta}_k)$ is:

$$\{Q'(\boldsymbol{\vartheta}_k)\}^t = \left(\frac{1 - P(D)}{B_k^0 B_k^1}, \frac{P(D)}{B_k^0 B_k^1}\right)$$

and $V(\widehat{\boldsymbol{\vartheta}}_k - \boldsymbol{\vartheta}_k)$ is the (2×2) diagonal matrix, with diagonal equal to

$$[N_0^{-1}P(T_k \mid D_0)\{1 - P(T_k \mid D_0)\},\ N_1^{-1}P(T_k \mid D)\{1 - P(T_k \mid D)\}].$$

A $100(1 - \alpha)\%$ valid Wald-type confidence interval for $\text{logit}\{P(T_k)\}$ is:

$$Q(\widehat{\boldsymbol{\vartheta}}_k) \pm z_{1-\alpha/2}\sqrt{\hat{V}\{Q(\widehat{\boldsymbol{\vartheta}}_k)\}}, \qquad k = 1, \ldots, K,$$

where $\hat{V}\{Q(\widehat{\boldsymbol{\vartheta}}_k)\}$ is obtained from substituting sample analogs for population parameters in $V\{Q(\widehat{\boldsymbol{\vartheta}}_k)\}$; the anti-logit transform yields the one for $P(T_k)$.

4.4 Adjustments for sparse false-positive counts

In settings that require a very high target specificity, for example in a screening population where a false-positive potentially induces patient harm, , the number of false-positive counts n_{0k} (i.e., controls misclassified into positive test category $T_k, k = 1, \ldots, K$) can be sparse, especially when the number of positive readout categories K is large relative to N_0. In these settings, because of the large weight $1 - P(D)$ assigned to $P(T_k \mid D_0)$ in the analytical form of PVP_k, PVN_k and the marginal test distribution, this sparseness can lead to instability in estimation and inference procedures for these metrics. Of particular concern are settings with $n_{0k} = 0$, referred to as “sampling zeros”, that result in an estimate of $P(T_k \mid D_0)$ that values 0.

Recall the sampling model for the control group test readout counts is multinomial with parameters N_0 and $\{P(T_k \mid D_0): k = 0, \ldots, K - 1\}$, which is a saturated model. Authors have advocated for adding ½ to each cell count for saturated models with sparse data in

order to reduce bias in resulting estimation (16,17,18). Using this approach, we obtain $n^*_{0k} = n_{0k} + \frac{1}{2}, k = 0, \ldots, K$. The resulting estimator for $P(T_k \mid D_0)$ is $\tilde{P}(T_k \mid D_0) = n^*_{0k}/\{N_0 + 0.5(K+1)\}$. The variance of the estimator $\tilde{P}(T_k \mid D_0)$ is obtained by multiplying the multinomial variance by $N_0^2/\{N_0 + 0.5(K+1)\}^2$, which approaches 1 as N_0 increases so that no adjustments are explicitly made to variance formulas for estimators of PVP_k, PVN_k and the marginal test distribution when implementing this adjustment. This adjustment may also be warranted for disease states when estimating $P(T_k \mid D_j)$ with sparse observed data and relatively high conditional incidence $P(D_j \mid D), j = 1, \ldots, J$.

5. Stratified analyses using case-control sampling

Analyses of subgroups of Table 2 arising from partitions defined by factors other than disease or test readout variables, e.g. demographic factors, proceed naturally using the corresponding methods above on the partitioned data.

For conditional specificity metrics given the stratum, the proposed estimation and inference methodology for the setting of a compound random variable in section 4.1 should be applied since the number of control samples in the stratum is in general a random variable.

Stratified analyses of predictive-value and marginal test distribution metrics require stratum-specific disease incidence values that may not be available from an external source; in such settings, conditional incidence estimates given disease state and stratum derived from the observed data under a random sampling model can be paired with an overall disease incidence value given the stratum. Furthermore, estimation and inference for predictive-value metrics needs to additionally accommodate the random number of control samples in the stratum; similarly for the marginal test distribution, where the total number of cases also needs to be considered as random.

When it is of interest to stratify analysis based explicitly on disease stage, two situations arise. Let the integer valued variable S denote disease stage. Intrinsic accuracy types of analyses proceed directly using the methodology above, where a conditional test readout distribution corresponding to a particular disease stage is analyzed using a Binomial model for the count of the number of cases of disease state D_k of the given stage $S = s$ with parameters $N_1$and $P(D_k, S = s \mid D)$.

In contrast, when investigating predictive value positive for a given test readout category, this metric can be decomposed into sub-parts corresponding to each disease stage. Define

$$PVP_k^{(S)} = \frac{P(T_k \mid D_k, D, S)P(S \mid D_k, D)P(D_k \mid D)P(D)}{P(T_k \mid D_0)\{1 - P(D)\} + P(D)\sum_{j=1}^{J} P(T_k \mid D_j)P(D_j \mid D)}, \quad k = 1, \ldots, K.$$

An estimator of $PVP_k^{(S)}$ is obtained analogously as for PVP_k with either the sample-based analog estimator for $P(\ S \mid D_k, D\ )$ or one sourced from a national registry or published large population study.

To conduct inference for $PVP_k^{(S)}$ at the stage value $S = s_0$, the logit transform is:

$$Z_{S=s_0}(\boldsymbol{\psi_k}) = \log\left\{\frac{P(T_k \mid D_k, S = s_0)P(S = s_0 \mid D_k)P(D_k \mid D)P(D)}{R_U(\boldsymbol{\varphi_k}) + P(T_k \mid D_k, S \neq s_0)P(S \neq s_0 \mid D_k)P(D_k \mid D)P(D)}\right\}, k = 1, \ldots, K,$$

where

$$\boldsymbol{\psi_k} = \{\boldsymbol{\varphi_k}, P(T_k \mid D_k, S = s_0), P(T_k \mid D_k, S \neq s_0), P(S = s_0 \mid D_k), P(S \neq s_0 \mid D_k)\}.$$

The delta-method may be used as shown above to obtain a variance estimator for $Z_{S=s_0}(\widehat{\boldsymbol{\psi}}_{\boldsymbol{k}})$, where $\widehat{\boldsymbol{\psi}}_{\boldsymbol{k}}$ is the sample analog of $\boldsymbol{\psi_k}$, from which an asymptotic $100(1-\alpha)\%$ Wald-type confidence interval for $\text{logit}\left\{PVP_k^{(S)}\right\}$ can analogously be constructed.

6. Simulation Study

A numerical study was conducted to investigate the performance of the proposed methodology in several simulated real-world scenarios. Two general clinical use-case settings were considered, corresponding to either a hypothesized screening or diagnostic test. The general format used *J=3* disease states and *K=2* positive test readouts as displayed below in Table 4.

Table 4:

The general sampling framework is described as follows. The number of controls, N_0 and the total number of cases, N_1, were fixed in advance. The number of enrolled cases of each disease state followed a multinomial sampling distribution with parameters N_1 and $(p_1, p_2, p_3 = 1 - p_1 - p_2)$. The value of the test was ascertained on each subject according to the following metrics:

Table 5:

Given a choice for population disease incidence, $P(D)$, values for predictive value positive, predictive value negative and the marginal test distribution corresponding to a setting of Table 5 can analytically be directly calculated.

First, a hypothesized cancer screening test setting presumed the goal was to maximize cancer detection at high specificity in order to minimize the potentially high-risk burden associated with downstream clinical workup of cancer-free individuals. The intended-use population was considered a high-risk population that would benefit from screening, so that $P(D) = 0.016$. For the hypothesized cancer diagnostic setting, the presumed goal was to minimize false negative results in an intended use population with suspicion of cancer, for example individuals presenting with clinical signs, symptoms and/or findings. Here, the cancer incidence was taken as $P(D) = 0.07$. For both settings, two choices for the distribution of case states among all N_1 cases were considered: $p_1 = P(D_1 \mid D) = 0.5$, $p_2 = P(D_2 \mid D) = 0.4$, and, $p_1 = P(D_1 \mid D) = 0.5$, $p_2 = P(D_2 \mid D) = 0.1$. Each simulation exercise generated 10,000 independent realizations from the given population setting.

Tables 6a, 6b and 6c present the mean of bias, the 95% confidence interval (CI) coverage indicator and the CI width for target parameters across simulation iterations in the screening setting under several sample sizes for the particular scenario $p_1 = P(D_1 \mid D) = 0.5$, $p_2 = P(D_2 \mid D) = 0.4$.

Table 6a:

Table 6b:

Table 6c:

Tables 7a, 7b and 7c present the mean of bias, the 95% CI coverage indicator and the CI width for target parameters in the diagnostic setting under several sample sizes for the particular scenario $p_1 = P(D_1 \mid D) = 0.5$, $p_2 = P(D_2 \mid D) = 0.4$.

Table 7a:

Table 7b:

Table 7c:

For both settings considered, coverage of the 95% CIs was sustained for all scenarios. Bias was very low for the diagnostic setting where the likelihood of a test positive readout was relatively large (about 27% for each positive test category). For the screening settings, the estimators of predictive value positive are based on very small fractions of the dataset (about 1.5% for each positive readout); the bias of these estimators decreased from about 2% to 1% to 0.5% as sample size increased. As

expected, the widths of all CIs became narrower with increasing sample size. Because the predictive value positive metrics were much smaller in the diagnostic setting, the CI widths were roughly 10-fold shorter compared to the screening setting.

When comparing results between scenarios with $p_2 = 0.4$ and $p_2 = 0.1$, because the value for PVP_2 is smaller when $p_2 = 0.1$, in turn, CI widths also decreased. For example, in the screening setting, for *N=500,* the CI width shrunk from about 30% with $p_2 = 0.4$ to about 12% with $p_2 = 0.1$.

When comparing settings that considered cancer incidence rates fixed, for example when obtained from a population registry, versus random, when estimating conditional incidence from the observed data, generally only minor differences were observed. The largest observed effect was in the setting with smaller magnitudes of predictive value positive (less than 15%), where the width of CIs when rates were considered fixed was roughly 0.5%, 0.4% and 0.3% shorter for *N=500, 1000, 2000*, respectively, for both screening and diagnostic settings.

Finally, a numerical study was conducted in a setting where specificity was very high so that an investigation of the proposed adjustment to the control group classification counts was warranted. The screening test setting described above was evaluated where specificity was increased from 0.98 to 0.995, with $p_1 = P(D_1 \mid D) = 0.5$, $p_2 = P(D_2 \mid D) = 0.1$. Table 8 shows results for predictive value positive and the marginal test distribution inference for three sample sizes. The expected false-positive count for each positive test category was 1.25, 2.5 and 5 for sample sizes 500, 1000 and 2000, respectively. Confidence interval coverage corresponding to inference on unadjusted data was observed to be very low for $N_0 = N_1 = 500$, however coverage did improve with increasing sample size. Similarly, bias of predictive value positive estimators on unadjusted data was large for $N_0 = N_1 = 500$ but did become smaller as sample size increased; both adjusted and unadjusted estimators for the marginal test distribution appeared unbiased. Inference corresponding to adjusted false-positive counts enjoyed great improvement in confidence interval coverage compared to the unadjusted counterparts, especially for the smallest sample size configuration. Furthermore, estimation with adjusted false-positive counts greatly reduced bias for PVP_2 compared to unadjusted data,

Based on this simulation result, a general rule-of-thumb for implementing the adjustment to control group classification counts suggests its use in settings where the expected false-positive count for one or more positive test readout categories is less than 5 subjects.

Table 8:

7. Data Analysis Example

The proposed MCED analysis framework was applied to a published real-world dataset (14). Here, the performance of a multi-cancer detection test using methylation signatures from cell-free DNA that provides cancer signal origin for positive tests was evaluated using a case-control design. The results reported for the Validation dataset are the focus of analysis here. This dataset consisted of $N_0 = 610$ non-cancer controls and $N_1 = 654$ cancer cases. Cancer incidence rates were obtained using the 2022 SEER database (19). The results presented in (14) show "crude sensitivity" performance of this test, that disregard incorrect CSO classification errors, by cancer stage in their Figure 5, in addition to predictions of CSO for multiple cancer types in their Figure 6. Because the implementation of the full suite of proposed methods described in our paper requires data on test Negative classifications for each cancer type, the analysis reported herein restricts attention to the cancer types reported in Figure 5 of (14). Also, because the data reported in Figure 6 B. of (14) did not directly permit construction of a two-way frequency distribution in the format of Table 2, the following two assumptions were necessary. There were 4 stomach cancer cases that received CSO classification, but the total number of such cases were unreported; we assumed there were 5 total stomach cancer cases with 1 false-negative prediction since overall crude sensitivity (disregarding CSO) was reported as 76.4% (p. 753 of (14), second sentence). There was 1 gallbladder case receiving CSO, the total number of such cases were not reported; we assumed there was 1 gallbladder case and no false-negative predictions for gallbladder.

Note that, although Figure 5 in (14) presents crude sensitivity by cancer type and stage, this metric only evaluated whether the subject was Negative or not on the test, without describing the CSO prediction accuracy per case type. That is, case types with a not Negative but incorrect CSO prediction were not considered as an error in the crude sensitivity calculation in Figure 5 of (14). Using our proposed methods, Table 9 below shows cancer-specific false-negative (1 – crude sensitivity) and intrinsic accuracy estimates by cancer type for all combined stages. Note that, unlike the results reported in Figure 5 of (14), the results in Table 9 below consider a not Negative test with incorrect CSO as a classification error, which is a more clinically relevant quantity. Because the Figure 6 B. of (14) is not stratified on cancer stage, we were not able to conduct a stage-stratified analysis.

Table 9:

In consideration of "TOO accuracy", an "overall accuracy" of 93% was reported in (14). This metric is not useful for assessing clinical validity of an MCED test for at least two

main reasons. First, this is an aggregate calculation, not providing information per-CSO readout category. Second, this calculation conditions on those cancer case types targeted by the test that had a positive (not Negative) readout; prospectively, a subject's cancer status and type is unknown at test issuance. Because this metric conditions on unknown information about the subject receiving the test as well as the test readout, no clear population analog is available to facilitate interpretation. Actionable and interpretable information for a practicing physician from such an MCED test will quantify in the intended use population the likelihood that a given CSO readout corresponds to a cancer case in the corresponding bodily site; this is measured by CSO-specific predictive value positive. Estimates and confidence intervals for predictive value positive and the marginal test distribution for the data reported in Figure 6 B. of (14) are provided below in Table 10. Because of the sparse false positive counts, the proposed correction that adds ½ to each non-cancer classification count was implemented.

Table 10:

Regarding the marginal test distribution, a Negative result is expected in 97.85% of all individuals from the target population receiving the test. Of the 2.15% of individuals receiving a not Negative test, 0.91% are expected to have an "Other" CSO readout; note this is more than 42% of all not Negative test readouts; thus, less than 58% of subjects who receive a not Negative readout will have a specific cancer signal of origin prediction. Predictive value negative was very high at 99.4%, as expected in a low disease incidence population. Values for predictive value positive ranged in magnitude from 7% for kidney to 52.6% for Lung, however, all 95% confidence intervals were very wide, resulting in much uncertainty in the clinical validity and potential contribution to clinical utility of this test base on the results of this dataset.

A graphical illustration that attempts to summarize the "cost-benefit" of an MCED test is shown below in Figure 1 for the test described in (14). Here, a "target region" is outlined as having a "benefit", that is, intrinsic accuracy for cancer detection, of at least 0.5 and a "cost", that is, incorrect CSO prediction, not more than 0.5. Based on this plot, it appears that the MCED test presented in (14) may have a desirable cost-benefit for Lung and CRC cancers, although the confidence intervals for predictive value positive of these two CSO readouts are wide.

Figure 1:

8. Discussion

The concept of testing for the presence of any one or more of a set of targeted disease types (such as multiple targeted cancer types using MCED tests) in a single blood draw is very attractive. These tests require unbiased estimation techniques and valid inference procedures for relevant performance metrics on a per-disease type level to properly quantify clinical validity that in-turn yields meaningful contributors to clinical utility assessments. Evaluating MCED tests using aggregate metrics across cancer types generally masks biological heterogeneity and cancer type-specific performance, thereby impeding a proper understanding of the test's relevant clinical validity. In the context of a case-control study, our approach presents analytical results for Classification-specific Predictive Performance (CSPP) that establish a framework for calculation of unbiased estimators for cancer-specific intrinsic accuracy, CSO-specific predictive value positive and the marginal test classification distribution, along with corresponding valid confidence interval estimation. The methodology has the potential to assess the value of an MCED test for informing diagnostic pathways in healthcare systems, facilitating MCED test optimization, standardizing comparisons of competing tests, aiding in the design of prospective studies, and meaningfully informing benefit-risk analyses and resulting follow-up decisions. Such capabilities can accelerate diagnosis, reduce healthcare costs, and enhance equitable access—attributes critical to regulatory and clinical adoption.

Conventional evaluations of MCED tests generally rely on three statistically disconnected metrics. One is an "overall PPV" that aggregates positive test results across all CSO readouts. This metric provides no information about the expected predictive ability of a given CSO readout; the probability of the correct case-type per given CSO readout is a crucial piece of information for informing clinical decision making. Another common metric is referred to as "TOO accuracy", that calculates the overall observed proportion of correct localization among those targeted cancer case types with a positive test readout. Because "TOO accuracy" conditions on both a positive test result and membership to a targeted cancer panel (unknown at test administration), there is no clear population analog to facilitate its interpretation. This fact coupled with "TOO accuracy" not directly incorporating information on disease incidence results in rendering a generally uninterpretable quantity for informing the test's clinical validity. A third misleading metric often used in the analysis of an MCED test that provides CSO, is a version of cancer-specific "sensitivity" that does not separate false-positive from true-positive test results. This aggregate and uninformative "crude sensitivity" metric does not quantify the test's true intrinsic accuracy per cancer type, rather, it only evaluates the test's ability to identify whether a cancer sample is Negative or not. These methodologically fragmented metrics often used in published

results for MCED tests obscure underlying heterogeneity in test performance, thereby precluding a properly informed evaluation of the test's clinical validity that impedes an understanding of the test's potential for aiding clinical decisions. Our proposed CSPP methodology establishes a framework that allows for unbiased estimates of cancer-specific intrinsic accuracy, CSO-specific predictive value positive and the marginal test classification distribution along with valid corresponding confidence intervals, thereby offering a clearly focused, clinically pointed and relevant lens on meaningful and actionable metrics for informing the risk-benefit profiles for each disease state and test readout category.

The shortcomings of conventional MCED evaluations become clearly evident when attempting to use corresponding information as a basis to actionably inform clinical decisions on a per-patient level. Overall PPV and "TOO accuracy" do not provide a practicing clinician, with a given CSO test readout in-hand for a patient under care, the information to evaluate the test's readout to aid in clinical decision making, thereby not contributing vital information necessary for prioritizing appropriate follow-up diagnostics. An overall PPV value may in fact be constructed using crude sensitivity metrics and arise as a composite combination of these complement false-negative rates and varying cancer incidences that in general can mask an underlying heterogeneous intrinsic accuracy profile. "TOO accuracy" is also non-informative for clinical patient management, not only because it is aggregate in nature but also because it conditions on membership to a targeted cancer panel and a not Negative test result. Since a prospective subject's cancer status (and cancer type if cancer) is unknown at test issuance, "TOO accuracy" does not convey prospective clinically actionable information for use in patient management.

The CSPP framework introduced here presents methodologies that allow for unbiased estimation and valid inference for disease state-specific intrinsic accuracy and CSO-specific predictive value—the probability that a given positive CSO readout corresponds to the target cancer type—into a unified MCED analysis framework. These metrics assess variation in underlying cancer-type specific performance otherwise obscured by conventional aggregate metric counterparts, offering a delineated probabilistic basis for assessing the risk-benefit profile per CSO readout. Our proposed framework for targeting clinically actionable information can be used to inform decisions about whether a given result suggests a strategy of immediate invasive diagnostics, or initial evaluations with diagnostic laboratory testing, or imaging for first-line workup. Unlike conventional approaches to MCED performance quantification that only provide fragmented, unfocused and distorted informational snapshots, our framework provides a cohesive, focused, and interpretable probabilistic basis for quantifying MCED performance that can directly inform clinical decision-making and the choice for an actionable follow-up strategy that aims to improve outcomes while minimizing harm.

The accuracy of a given CSO readout can vary between early- and late-stage cancers due to biological differences in tumor burden or marker expression. Conventional metrics that report aggregate predictive value performance obscure the underlying variation corresponding to stage-specific heterogeneity. For cancers with established screening programs (breast, colorectal, cervical, lung), it is critical to obtain early-stage intrinsic accuracy to inform the likelihood of timely detection and the impact on altering prognosis. A test with an apparently favorable observed overall PPV might perform well in assigning a positive test readout to a common late-stage cancer (e.g. colorectal), but falter in early stages and/or capability for providing the correct CSO readout beyond not Negative, where routine screening modalities compete, rendering this aggregate metric misleading. Our methodology embeds capability for estimation and inference for stage-stratified intrinsic accuracy metrics per cancer type and furthermore breaks down CSO-specific predictive value positive into components corresponding to disease stage within a unified framework. Our proposed tools for MCED analysis can thus delineate performance gradients—e.g., an overall high colorectal CSO-PVP of say 0.85 may, for example, be decomposed into 0.35 for stages I and II and 0.50 in stages III and IV—offering a transparent basis for evaluating efficacy across the disease continuum and ensuring that screening-relevant cancers are not underserved by inflated, crude and stage-agnostic claims.

The capability of our proposed methodology to perform unbiased and valid inference for intrinsic accuracy metrics per cancer type and stage establishes a framework that can be leveraged to approximate the quantification of an MCED test's clinical validity in an asymptomatic screening intended-use (IU) population when using data from a case-control study. Because the cancer stage and/or type distribution observed in a case-control study may differ from that in the asymptomatic screening IU population (9, 10) during the sojourn time for preclinical cancer detection, estimation and inference for intrinsic accuracy metrics per cancer type (and overall) can be calibrated to the expected cancer incidence stage distribution per type (and further the cancer type incidence for overall intrinsic accuracy) that is expected in the asymptomatic screening IU population. Similarly, estimation and inference for CSO-specific predictive value positive can be calibrated to the expected incidence distribution in the IU population per cancer type and stage. The utilized incidence distribution per cancer type and stage can be varied in a sensitivity analysis if uncertainty in these values is present. In addition, there is the potential for spectrum bias in the control group of the case-control study; for example, if only healthy volunteers were recruited, the underlying incidence of precancerous and benign lesions in this group may not be representative of the non-cancer subjects in the asymptomatic IU screening population (9). If the investigational setting recorded information on such comorbidities, test performance could be stratified on these factors, and their contribution to estimation and inference for aggregate

specificity, CSO-specific predictive value positive and the marginal test distribution adapted to their expected incidence in the IU population in an attempt to remedy this source of potential bias.

The impact on clinical practice afforded by information obtained by the appropriate analysis of MCED tests extends beyond initial screening to include the potential for influencing subsequent testing strategies, where diagnostic intervention harms must be weighed against benefits. Conventionally reported aggregate metrics do not provide information for aiding in establishing a risk-benefit guided strategy that evaluates different choices for follow-up pathways, leaving clinicians without probabilistic anchors to enable objectively informed decisions that attempt to optimize outcomes and minimize harm for the individual patient. Our MCED analysis framework provides a basis for analysis that contributes relevant information to help form strategies that aim to optimize patient outcomes by aligning intervention intensity with the quantified confidence level of the reported test classification. Cancer signals resulting in a test readout category with an acceptably large corresponding predictive value positive may warrant immediate invasive diagnostics; whereas cancer signals corresponding to a test readout with a considered small predictive value positive suggest a stepwise, less invasive approach to mitigate procedural risks. For example, a colorectal cancer signal with a CSO-specific predictive value positive of 0.85 and 0.60 early-stage intrinsic accuracy may suggest an accelerated colonoscopy scheduling—balancing a 0.15 incorrect CSO readout risk against delayed diagnosis morbidity—whereas a pancreatic signal with a predictive value positive of 0.30 might first trigger imaging, reserving biopsy for confirmed findings. This stratification, unattainable with dissociated and aggregate metrics, informs clinical actionability by quantifying diagnostic yield against procedural hazards.

Another critical component of a test's performance is quantification of the cancer-specific false-negative and incorrect CSO classification proportions. A false-negative test result engenders false reassurance, which is particularly concerning for cancers with established screening guidelines; traditional metrics that fail to highlight this aspect may thus fail to signal when a negative result might undermine trust in routine screening programs. Incorrect CSO readout classifications can suggest unnecessary and potentially harmful intervention. Conventional metrics that calculate cancer-specific crude sensitivity simply assess the likelihood of a "positive test" and ignore the actual CSO classification. Such metrics thereby mask CSO classification errors and quantify an agnostic metric of the MCED test's ability to detect the cancer type under consideration. Our MCED test analysis framework's ability to both quantify cancer-specific false-negative proportions and intrinsic accuracy measures (for the correct corresponding CSO readout) directly address the pitfalls of conventional analysis

approaches by quantifying and decomposing detection failures across tumor types and stages. For example, a breast cancer false-negative proportion of 0.50 in early stages versus 0.20 in late stages might indicate a 50% miss rate for mammography-eligible lesions, risking patient complacency if a negative MCED test result is misinterpreted as a definitive finding. Furthermore, a reported positive test proportion of 80% for colorectal cancer may in fact include 30% of cases that incorrectly obtained an Upper GI or Lung CSO readout; the correct intrinsic accuracy is thus 50%. The transparency and focus underlying our analysis tools enables evaluation of an MCED test's compatibility with existing screening paradigms and provides a focused evidence generation that attempts to mitigate false reassurance by identifying cancer types where supplementary screening remains essential despite a Negative test readout.

Our proposed analytical framework is flexible in that it allows for a potentially large number of distinct disease states (J) in the IU population, not all of which that may have a corresponding CSO test readout category. The CSO capability and/or target cancer panel of the test (with K specific CSO readout categories) may not encompass all J disease states ($K < J$). Here, intrinsic accuracy metrics and CSO-specific predictive value positive are not defined for the disease states D_{K+1} to D_J, however, false-negative rates for these disease states can be analyzed by the proposed methodology. Another alternative test configuration here can include with the K positive readout categories an "Other" CSO readout category that potentially draws case states D_{K+1} to D_J. The main benefit of this approach is a potential reduction in distortion to CSO-specific predictive value positive of readout categories 1,...,K due to mis-classification of subjects with disease states D_{K+1} to D_J.

The analytical setting considered in this manuscript is for one and only one CSO prediction per individual. However, an alternative test configuration can report a "primary" CSO as well as a "secondary" CSO prediction for each individual. Although this configuration that presents the "top two" CSO predictions provides less precise clinical information, the predictive value positive for a pair of CSO predictions enjoys a larger target sample space and a potential for achieving greater numerical values. Extending the proposed analytical capabilities to accommodate this setting is currently underway.

In summary, the proposed Classification-specific Predictive Performance (CSPP) methodology is a unified analytical framework offering multiple key diagnostic metrics for quantifying the clinical validity of a multi-category test with multiple disease states: cancer state-specific intrinsic accuracy, cancer signal-of origin-specific predictive value positive , and marginal test classification distributions. By bringing these metrics together into a coherent statistical model for individual classification-specific (or

disease-specific) diagnostic accuracy, it addresses critical limitations inherent in conventional evaluation methods, that obscure classification-specific or disease-specific performance variations, resulting in diagnostic biases and misinformation arising from use of aggregate and unfocused metrics.

Adopting the proposed test analysis framework when quantifying clinical validity can facilitate the clinical interpretability of complex, multi-category and/or multi-cancer diagnostic data, providing a clear and actionable probabilistic basis to support decision-making. It facilitates informed risk-benefit assessments across disease (cancer) types and stages, thereby aiding decisions on follow-up clinical strategies ranging from intervention to monitoring. Furthermore, CSPP's detailed insights can inform future study design and sizing decisions for multi-disease diagnostics, while also standardizing comparisons across competing tests and diverse diagnostic platforms, whether biological assays, digital AI-based methods, or hybrid approaches. By attempting to integrate disease-specific diagnostic capability into a single, interpretable, unbiased and valid analytical structure, this methodology carries the potential to streamline diagnostic decision-making for clinicians, policymakers, scientists, and patients alike, promoting optimized, equitable, and evidence-based implementation of multi-disease tests in healthcare.

Acknowledgement

The authors thank a reviewer whose insightful and detailed comments improved the contextualization and presentation of the manuscript.

References

1. Health Quality Ontario. Pharmacogenomic Testing for Psychotropic Medication Selection: A Systematic Review of the Assurex GeneSight Psychotropic Test. *Ont Health Technol Assess Ser.* 2017; 17:1-39.
2. Nix P, Mundt E, Manley S, Coffee B, Roa B. Functional RNA Studies Are a Useful Tool in Variant Classification but Must Be Used With Caution: A Case Study of One *BRCA2* Variant. *JCO Precis Oncol.* 2020; 4:730-735.
3. Ahlquist DA. Universal cancer screening: revolutionary, rational, and realizable. *npj Precision Onc.* 2018; 2:1-5.
4. Kang SK, Gulati R, Moise N, Hur C, Elkin EB. Multi-Cancer Early Detection Tests: State of the Art and Implications for Radiologists. *Radiology.* 2025; 314:e233448.
5. LeeVan E, Pinsky P. Predictive Performance of Cell-Free Nucleic Acid-Based Multi-Cancer Early Detection Tests: A Systematic Review. *Clin Chem.* 2024; 70:90-101.

6. Rubin R. Questions Swirl Around Screening for Multiple Cancers with a Single Blood Test. *JAMA.* 2024; 331:1077-1080.
7. Rubinstein WS, Patriotis C, Dickherber A, Han PKJ, Katki HA, LeeVan E, Pinsky PF, Prorok PC, Skarlupka AL, Temkin SM, Castle PE, Minasian LM. Cancer screening with multicancer detection tests: A translational science review. *CA Cancer J Clin.* 2024; 74:368-382.
8. Hoffman RM, Wolf AMD, Raoof S, Guerra CE, Church TR, Elkin EB, Etzioni RD, Shih YT, Skates SJ, Manassaram-Baptiste D, Smith RA. Multicancer early detection testing: Guidance for primary care discussions with patients. *Cancer.* 2025; 131:e35823.
9. Putcha G, Gutierrez A, Skates S. Multicancer Screening: One size does not fit all. *JCO Precision Oncology.* 2021; 5:574-576.
10. Etzioni R, Gulati R, Patriotis C, Rutter C, Zheng Y, Srivastava S, Feng Z. Revisiting the standard blueprint for biomarker development to address emerging cancer early detection technologies. *J Natl Cancer Inst.* 2024; 116:189-193.
11. Zhao Y, Gulati R, Lange J, Olivas-Martinez A, Raoof S, Zheng Y, Feng Z, Etzioni R. Sensitivity Measures in Studies of Cancer Early Detection Biomarkers. *Cancer Epidemiol Biomarkers Prev.* 2025; 34:944-951.
12. Pinsky P, Lange J, Etzioni R. Estimating stage-specific sensitivity for cancer screening tests. *J Med Screen.* 2023; 30:69-73.
13. Mercaldo N, Lau K, Zhau Z. Confidence intervals for predictive values with an emphasis to case-control studies. *Statistics in Medicine.* 2007; 26:2170-2183.
14. Liu M, Oxnard G, Klein E, Swanton C, Seiden M, & on behalf of the CCGA Consortium. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Annals of Oncology.* 2020; 31:745-759.
15. Agresti A, Gottard A. Comment: Randomized confidence intervals and the Mid-*P* approach. *Statistical Science.* 2005; 20:367-371.
16. Goodman LA. Interactions in multi-dimensional contingency tables. *Annals of Mathematical Statistics.* 1964; 35:632-646.
17. Goodman LA. The multivariate analysis of qualitative data: Interaction among multiple classifications. *Journal of the American Statistical Association.* 1970; 65: 226-256.
18. Goodman LA. The analysis of multidimensional contingency tables: Stepwise procedures and direct estimation methods for building models for multiple classifications. *Technometrics.* 1971; 13:33-61.
19. SEER*Explorer: An interactive website for SEER cancer statistics [Internet]. Surveillance Research Program, National Cancer Institute; 2024 Apr 17. Available from: https://seer.cancer.gov/statistics-network/explorer/. Data

source(s): SEER Incidence Data, November 2023 Submission (1975-2021), SEER 22 registries.

Tables

	Test -	Test +	TOTAL
Control	n_{00}	n_{01}	N_0
Case	n_{10}	n_{11}	N_1
TOTAL	n_{+0}	n_{+1}	N

Table 1: Standard two-way frequency distribution of test versus disease state.

	Test=0	***Test=1***	⋯	***Test=K***	**TOTAL**
$\boldsymbol{D_0}$	n_{00}	n_{01}	⋯	n_{0K}	N_0
$\boldsymbol{D_1}$	n_{10}	n_{11}	⋯	n_{1K}	n_{1+}
⋮	⋮	⋮	⋯	⋮	⋮
$\boldsymbol{D_K}$	n_{K0}	n_{K1}	⋯	n_{KK}	n_{K+}
⋮	⋮	⋮	⋯	⋮	⋮
$\boldsymbol{D_J}$	n_{J0}	n_{J1}	⋯	n_{JK}	n_{J+}
TOTAL	n_{+0}	n_{+1}	⋯	n_{+K}	N

Table 2: General two-way frequency distribution of a multi-category test and disease states.

	Negative	Uterus	UGI	Prostate	PG	Lung	HN	CRC	Breast	Kidney	Others	TOTAL
Control	606	0	0	0	0	0	0	0	0	0	4	610
Uterus	27	8	0	0	0	0	0	0	0	0	1	36
UGI	5	0	17	0	1	2	0	1	0	0	0	26
Prostate	74	0	0	10	0	0	0	0	0	0	0	84
PG	6	0	1	0	30	0	0	0	1	0	2	40
Lung	39	0	0	0	0	71	1	0	0	0	0	111
HN	8	0	0	0	0	1	15	0	1	0	0	25
CRC	12	0	2	0	0	1	0	38	0	0	0	53
Breast	63	0	0	0	0	0	0	0	40	0	1	104
Kidney	22	0	0	0	0	0	0	0	0	3	0	25
Others	50	0	2	0	0	4	2	0	1	0	91	150
TOTAL	912	8	22	10	31	79	18	39	43	3	99	1264

Figure 6 B. from Liu et al. (2020) (with modification): modification as described in Section 7 to permit analysis of our proposed suite of performance metrics. UGI: upper gastrointestinal; PG: Pancreas and gallbladder; HN: Head and neck; CRC: colon and rectum.

	Test=0	***Test=1***	⋯	***Test=K***	**TOTAL**
Controls$(\boldsymbol{D_0})$	n_{00}	n_{01}	⋯	n_{0K}	N_0
Cases $(\boldsymbol{D})$	$n_{+0} - n_{00}$	$n_{+1} - n_{01}$	⋯	$n_{+K} - n_{0K}$	N_1
TOTAL	n_{+0}	n_{+1}	⋯	n_{+K}	N

Table 3: Collapsed version of Table 2 to facilitate inference for marginal Test category distribution.

	Test=0 $(\boldsymbol{T_0})$	***Test=1*** $(\boldsymbol{T_1})$	***Test=2*** $(\boldsymbol{T_2})$	**TOTAL**
$\boldsymbol{D_0}$	$E(n_{00}) = N_0 SP_0$	$E(n_{01}) = N_0 \beta_1$	$n_{02} = N_0 - n_{00} - n_{01}$	N_0
$\boldsymbol{D_1}$	$E(n_{10}) = n_{1+}\alpha_1$	$E(n_{11}) = n_{1+}SE_1$	$n_{12} = n_{1+} - n_{10} - n_{11}$	$E(n_{1+}) = N_1 p_1$
$\boldsymbol{D_2}$	$E(n_{20}) = n_{2+}\alpha_2$	$n_{21} = n_{2+} - n_{20} - n_{22}$	$E(n_{22}) = n_{2+}SE_2$	$E(n_{2+}) = N_1 p_2$
$\boldsymbol{D_3}$	$E(n_{30}) = n_{3+}\alpha_{31}$	$E(n_{31}) = n_{3+}\alpha_{32}$	$n_{32} = n_{3+} - n_{30} - n_{31}$	$n_{3+} = N_1 - n_{1+} - n_{2+}$
TOTAL	n_{+0}	n_{+1}	n_{+2}	$N = N_0 + N_1$

Table 4: Two-way frequency distribution of disease and test for simulation study.

Notation	SP_0	β_1	α_1	A_1	α_2	A_2	α_{31}	α_{32}
Metric	$P(T_0 \mid D_0)$	$P(T_1 \mid D_0)$	$P(T_0 \mid D_1)$	$P(T_1 \mid D_1)$	$P(T_0 \mid D_2)$	$P(T_2 \mid D_2)$	$P(T_0 \mid D_3)$	$P(T_1 \mid D_3)$
Screen	0.98	0.01	0.25	0.65	0.30	0.50	0.60	0.20
Diagnostic	0.50	0.25	0.03	0.95	0.05	0.92	0.10	0.40

Table 5: Conditional test distribution given disease state for screening, diagnostic settings.

	Value	**Bias**	**Coverage**	**Width**
$\boldsymbol{SP_0}$	98	0	94.60	2.438
$\boldsymbol{A_1}$	65	0.006	95.35	11.774
$\boldsymbol{A_2}$	50	-0.023	95.30	13.772
$\boldsymbol{PVN_0}$	99.50	0	95.04	0.134
$\boldsymbol{PVP_1}$	31.25	2.35	93.32	32.794
$\boldsymbol{PVP_2}$	22.60	2.531	94.36	30.072
$\boldsymbol{P(T_0)}$	96.92	0	94.68	2.438
$\boldsymbol{P(T_1)}$	1.66	0	93.44	1.74
$\boldsymbol{P(T_2)}$	1.42	0	93.08	1.766

Table 6a: Screening simulation results for $N_0 = N_1 = 500$ with $P(\, D_2 \mid D \,) = 0.4$; all numbers are percents.

	Value	Bias	Coverage	Width
SP_0	98	0.002	93.70	1.695
A_1	65	-0.035	95.32	8.347
A_2	50	-0.002	94.92	9.767
PVN_0	99.50	0	94.71	0.094
PVP_1	31.25	1.15	94.54	23.448
PVP_2	22.60	1.209	94.89	20.939
$P(T_0)$	96.92	0.003	93.74	1.717
$P(T_1)$	1.66	-0.001	94.22	1.225
$P(T_2)$	1.42	-0.001	94.37	1.234

Table 6b: Screening simulation results for $N_0 = N_1 = 1000$ with $P(D_2 \mid D) = 0.4$; all numbers are percents.

	Value	Bias	Coverage	Width
SP_0	98	0.003	94.49	1.18
A_1	65	0.002	95.12	5.906
A_2	50	-0.019	94.76	6.918
PVN_0	99.50	0	95.09	0.067
PVP_1	31.25	0.513	94.61	16.494
PVP_2	22.60	0.629	95.23	14.506
$P(T_0)$	96.92	0.003	94.77	1.211
$P(T_1)$	1.66	0.002	94.31	0.865
$P(T_2)$	1.42	-0.005	94.62	0.865

Table 6c: Screening simulation results for $N_0 = N_1 = 2000$ with $P(D_2 \mid D) = 0.4$; all numbers are percents.

	Value	Bias	Coverage	Width
SP_0	50	0.005	93.94	8.689
A_1	95	0	95.79	5.572
A_2	92	0.021	95.55	7.657
PVN_0	99.33	0.001	95.08	0.567
PVP_1	12.34	0.076	95.06	3.859
PVP_2	9.82	0.033	95.42	3.375
$P(T_0)$	46.81	0.003	94.36	8.129
$P(T_1)$	26.94	-0.032	95.11	7.066
$P(T_2)$	26.25	0.029	94.88	7.073

Table 7a: Diagnostic simulation results for $N_0 = N_1 = 500$ with $P(D_2 \mid D) = 0.4$; all numbers are percents.

	Value	Bias	Coverage	Width
SP_0	50	0.011	94.83	6.143
A_1	95	0.009	94.98	3.875
A_2	92	-0.019	95.13	5.373
PVN_0	99.33	0	94.91	0.396
PVP_1	12.34	0.034	95.15	2.713
PVP_2	9.82	0.019	94.83	2.378
$P(T_0)$	46.81	0.01	95.11	5.758
$P(T_1)$	26.94	-0.01	94.97	5.004
$P(T_2)$	26.25	0	94.80	5.005

Table 7b: Diagnostic simulation results for $N_0 = N_1 = 1000$ with $P(D_2 \mid D) = 0.4$; all numbers are percents.

	Value	Bias	Coverage	Width
SP_0	50	-0.017	94.86	4.33
A_1	95	-0.01	95.23	2.725
A_2	92	0.007	94.91	3.775
PVN_0	99.33	-0.001	95.20	0.278
PVP_1	12.34	0.012	95.11	1.913
PVP_2	9.82	0.009	95.10	1.676
$P(T_0)$	46.81	-0.016	94.98	4.075
$P(T_1)$	26.94	-0.003	94.98	3.541
$P(T_2)$	26.25	0.019	95.19	3.542

Table 7c: Diagnostic simulation results for $N_0 = N_1 = 2000$ with $P(D_2 \mid D) = 0.4$; all numbers are percents.

	Value	Bias	Coverage	Width
$N_0 = N_1 = 500$				
$\boldsymbol{PVP_1}$	56.156	-3.023 (2.834)	95.51 (68.98)	44.75 (36.26)
$\boldsymbol{PVP_2}$	14.981	-0.813 (2.563)	95.84 (73.23)	24.187 (22.46)
$\boldsymbol{P(T_0)}$	98.54	-0.191 (-0.004)	95.94 (90.81)	1.44 (1.18)
$\boldsymbol{P(T_1)}$	0.926	0.093 (0.002)	97.61 (70.77)	1.006 (0.78)
$\boldsymbol{P(T_2)}$	0.534	0.098 (0.002)	96.13 (69.45)	1.066 (0.82)
$N_0 = N_1 = 1000$				
$\boldsymbol{PVP_1}$	56.156	-1.576 (1.316)	92.74 (89.64)	33.65 (32.139)
$\boldsymbol{PVP_2}$	14.981	-0.283 (1.352)	93.4 (89.21)	17.635(18.075)
$\boldsymbol{P(T_0)}$	98.54	-0.098 (-0.002)	95.4 (92.66)	0.942 (0.855)
$\boldsymbol{P(T_1)}$	0.926	0.049 (0.004)	90.7 (90.93)	0.662 (0.594)
$\boldsymbol{P(T_2)}$	0.534	0.049 (-0.002)	89.51 (90.09)	0.678 (0.601)
$N_0 = N_1 = 2000$				
$\boldsymbol{PVP_1}$	56.156	-0.647 (0.907)	93.43 (93.50)	24.99 (24.846)
$\boldsymbol{PVP_2}$	14.981	-0.107 (0.680)	94.44 (92.87)	12.82 (13.261)
$\boldsymbol{P(T_0)}$	98.54	-0.047 (0.004)	94.69 (93.09)	0.638 (0.606)
$\boldsymbol{P(T_1)}$	0.926	0.023 (-0.002)	94.40 (91.60)	0.45 (0.427)
$\boldsymbol{P(T_2)}$	0.534	0.024 (-0.002)	93.29 (93.51)	0.456 (0.432)

Table 8: Performance with adjusted control group classification counts in screening setting with $SP_0 = 0.995$, $P(D_2 \mid D) = 0.1$ and various sample sizes; results for original (unadjusted) data shown in parentheses; all numbers are percents.

	False Negative	Intrinsic Accuracy IA(k)	IA(k): CI-low	IA(k): CI-up
Uterus	75.0	22.2	11.4	38.8
Upper GI	19.2	65.4	45.3	81.2
Prostate	88.1	11.9	6.50	20.8
Pancreas & Gallbladder	15.0	75.0	59.2	86.1
Lung	35.1	64.0	54.6	72.4
Head & Neck	32.0	60.0	39.9	77.2
CRC	22.6	71.7	58.1	82.2
Breast	60.6	38.5	29.6	48.2
Kidney	88.0	12.0	3.80	31.8
Others	33.3	60.7	52.6	68.2

Table 9: Empirical false-negative proportion and intrinsic accuracy with 95% confidence interval (CI) for 9 cancer types and "Others". All numbers are percents. CRC: colon and rectum; GI: gastrointestinal.

	P(T=k)	P(T=k): CI-low	P(T=k): CI-up	PVP(k)	PVP(k): CI-low	PVP(k): CI-up
Non-Cancer	97.85	96.68	98.61	99.4*	99.3*	99.4*
Uterus	0.10	0.01	0.96	10.4	0.6	67.6
Upper GI	0.12	0.02	0.74	20.4	2.1	74.9
Prostate	0.10	0.01	0.91	23.4	1.8	83.7
Pancreas & Gallbladder	0.14	0.03	0.68	31.3	2.9	87.5
Lung	0.24	0.09	0.61	52.6	10.4	91.4
Head & Neck	0.12	0.02	0.78	23.1	2.3	79.6
CRC	0.16	0.04	0.65	49.1	5.9	93.7
Breast	0.17	0.04	0.63	43.6	5.5	91.2
Kidney	0.09	0.01	1.12	7	0.4	59.7
Others	0.91	0.44	1.89	27.1	12.9	48.3

Table 10: Estimated marginal test readout distribution and predictive value positive with 95% confidence interval (CI) for 9 cancer types and “Others”. Predictive value negative indicated with an “*”. All numbers are percents. CRC: colon and rectum; GI: gastrointestinal.

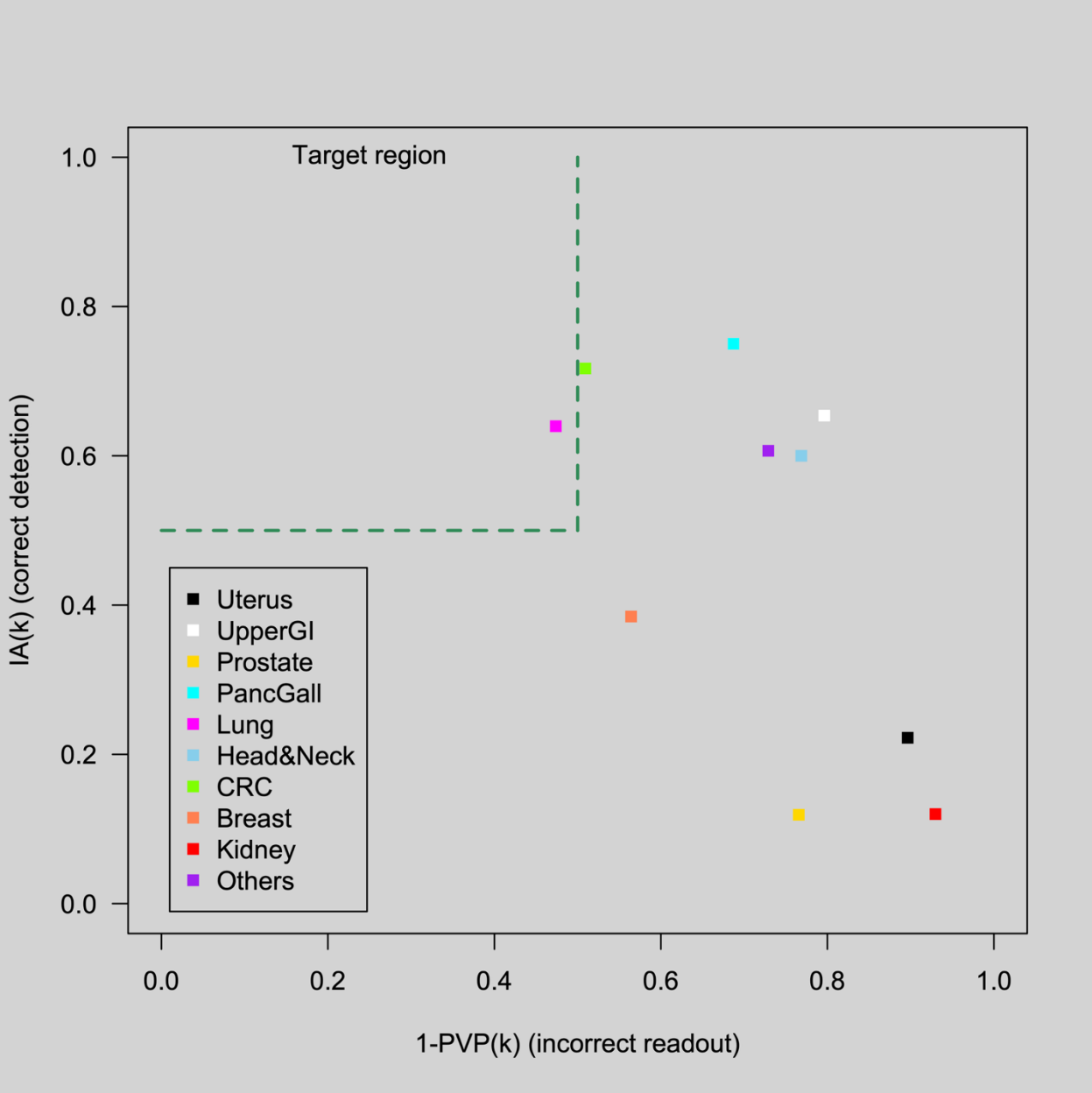

Figure 1: A “cost benefit” illustration of the MCED test presented in Liu et al. (14). The “cost” on the horizontal axis is the estimated probability of an incorrect call per CSO readout category; the “benefit” on the vertical axis is the estimated probability of a correct CSO call per cancer group.

Appendix

$P\left(n_{j+} > 0\right)$ is considered fixed when constructing variance estimators.

A1: Components of $\boldsymbol{U'(\varphi_k)}$, $k = 1, \dots, K$.

(1) $\frac{\partial U(\boldsymbol{\varphi_k})}{\partial P(T_k|D_0)} = -\{1 - P(D)\}/R_U(\varphi_k)$

(2) $\frac{\partial U(\boldsymbol{\varphi_k})}{\partial P(T_k|D_k)} = 1/P(T_k \mid D_k)$

(3) For $j = 1, \dots, J, j \neq k$:

$\frac{\partial U(\boldsymbol{\varphi_k})}{\partial P(T_k|D_j)} = -\{I(j \neq k)P\left(D_j \mid D\right)P(D)\}/R_U(\varphi_k)$

(4) For $k < J$:

$\frac{\partial U(\boldsymbol{\varphi_k})}{\partial P(D_k|D)} = \frac{1}{P(D_k|D)} + \frac{I(k \neq J)P(D)P(T_k|D_J)}{R_U(\varphi_k)}$

(5) For $k < J, j = 1, \dots, J-1, j \neq k$:

$\frac{\partial U(\boldsymbol{\varphi_k})}{\partial P(D_j|D)} = -P(D)\left\{I(k \neq j)P\left(T_k \mid D_j\right) - I(k \neq J)P\left(T_k \mid D_J\right)\right\}/R_U(\boldsymbol{\varphi_k})$

(6) For $k = J, j = 1, \dots, J-1$:

$\frac{\partial U(\boldsymbol{\varphi_k})}{\partial P(D_j|D)} = \frac{-1}{\left\{1 - \sum_{\ell=1}^{J-1} P(D_\ell|D)\right\}} - \frac{I(k \neq j)P(D)P(T_k|D_j)}{R_U(\varphi_k)}$

A2: Components of $\boldsymbol{W'(\varphi_k)}, k = 0, \dots, K$.

Define $R_W(\boldsymbol{\varphi_k}) = P(D)\sum_{j=1}^{J-1} P(T_k \mid D_j)P(D_j \mid D) + P(D)P(T_k \mid D_J)\left\{1 - \sum_{j=1}^{J-1} P\left(D_j \mid D\right)\right\}$

(1) $\frac{\partial W(\boldsymbol{\varphi_k})}{\partial P(T_k|D_0)} = 1/P(T_k \mid D_0)$

(2) For $j = 1, \dots, J-1$:

$\frac{\partial W(\boldsymbol{\varphi_k})}{\partial P(T_k|D_j)} = -P(D)P(D_j \mid D)/R_W(\boldsymbol{\varphi_k})$

(3) For $j = J$:

$\frac{\partial W(\boldsymbol{\varphi_k})}{\partial P(T_k|D_J)} = -P(D)\{1 - \sum_{\ell=1}^{J-1} P(D_\ell \mid D)\}/R_W(\boldsymbol{\varphi_k})$

(4) For $j = 1, \dots, J-1$:

$\frac{\partial W(\boldsymbol{\varphi_k})}{\partial P(D_j|D)} = -P(D)\{P(T_k \mid D_j) - P(T_k \mid D_J)\}/R_W(\boldsymbol{\varphi_k})$

A3: Components of $(2J \times 2J)$ matrix $\boldsymbol{V}(\widehat{\boldsymbol{\varphi}}_{\boldsymbol{k}} - \boldsymbol{\varphi_k}), k = 0, \dots, K.$

$$\boldsymbol{V}(\widehat{\boldsymbol{\varphi}}_{\boldsymbol{k}} - \boldsymbol{\varphi_k}) = \begin{pmatrix} N_0^{-1} P(T_k \mid D_0)\{1 - P(T_k \mid D_0)\} & 0 & \cdots & 0 \\ 0 & & \boldsymbol{V}_1 & \boldsymbol{V}_2 \\ \vdots & & & \\ 0 & & \boldsymbol{V}_2^t & \boldsymbol{V}_3 \end{pmatrix}$$

$\boldsymbol{V}_1 = \text{diag}(\sigma_{1k}^2, \cdots, \sigma_{Jk}^2)$, a$(J \times J)$ matrix, since $Cov\left(\frac{\tilde{A}_{jk}}{P(n_{j+}>0)}, \frac{\tilde{A}_{\ell k}}{P(n_{\ell+}>0)}\right) = 0, (j \neq \ell)$, under the assumptions:

(1) $E\left(Y_i^{(jk)} \mid Y_i^{(\ell k)}, n_{j+}, n_{\ell+}\right) = E(Y_i^{(jk)})$ and $E\left(Y_i^{(jk)} \mid n_{j+}, n_{\ell+}\right) = E(Y_i^{(jk)})$ these follow from the definition of $\tilde{A}_{jk}$ as a compound random variable.

(2) $P\{(n_{j+} > 0) \mid (n_{\ell+} > 0)\} = P\{(n_{j+} > 0)\}$, this is a reasonable assumption when similar disease types are grouped when defining $D_1 \cdots D_J$.

$$\boldsymbol{V}_2 = \begin{pmatrix} \eta_{11}^{(k)} & 0 & \dots & 0 \\ 0 & & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & \eta_{(J-1)(J-1)}^{(k)} \\ 0 & & \cdots & 0 \end{pmatrix}, \text{ a } (J \times (J-1)) \text{ matrix}$$

$\eta_{j\ell}^{(k)} = Cov\left(\frac{\tilde{A}_{jk}}{P(n_{j+} > 0)}, \frac{n_{\ell+}}{N_1}\right), j = 1, \dots, J, \ell = 1, \dots, J-1.$

Under the assumption $E\left(Y_i^{(jk)} \mid n_{j+}, n_{\ell+}\right) = E(Y_i^{(jk)})$,

when $j = \ell$:

$\eta_{j\ell}^{(k)} = P(T_k \mid D_j)\left\{\frac{1}{N_1} E(n_{j+} \mid n_{j+} > 0) - P(D_j \mid D)\right\}$

when $j \neq \ell$:

$\eta_{j\ell}^{(k)} = P\big(T_k \mid D_j\big)\left\{\frac{1}{N_1}E(n_{\ell+} \mid n_{j+} > 0) - P(D_\ell \mid D)\right\}$

= 0 when $E\big(n_{\ell+} \mid n_{j+} > 0\big) = E(n_{\ell+})$.

$\boldsymbol{V}_3$ is the $(J-1)\times(J-1)$ multinomial variance covariance matrix. Define the column vector $\boldsymbol{v} = \big\{P(D_1 \mid D), \ldots, P\big(D_{J-1} \mid D\big)\big\}^t$

$\boldsymbol{V}_3 = N_1^{-1}\{\mathrm{diag}(\boldsymbol{v}) - \boldsymbol{v}\boldsymbol{v}^{\boldsymbol{t}}\}$.