Research Article
Volume 13:1, 2022

Journal of Biometrics & Biostatistics

ISSN: 2155-6180 **Open Access**

Estimating and Comparing Diagnostic Laboratory Performance with Weighted Estimating Equations

Patricia Cooper Barfoot*, Stefan H. Steiner and R. Jock MacKay

Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Canada

Abstract

Patient test outcomes from diagnostic testing laboratories can be indicators of laboratory performance. One method of proficiency testing compares test results across laboratories to flag non-compliant laboratories. Under this type of proficiency test, the statistical approaches for estimating the test result and comparing these across laboratories have important implications. The proficiency test of fecal occult blood testing laboratories in Ontario compare estimates of pass rate for a particular test by laboratory based on data observed in the present (latest) month. Estimates by laboratory are compared to an acceptance interval determined by data across all laboratories. A laboratory is classified as non-compliant when its monthly rate is outside the acceptance interval. We note that monthly sample sizes have an important impact on the probability that a laboratory is classified in error and differences in the number of tests conducted across laboratories should be considered. We present an alternative approach (Weighted Estimating Equations, or WEE) for combining historical data to improve the precision of the estimate of present performance in the case that performance changes slowly over time. The WEE approach uses all available historical data through estimating functions that down-weight past data. We compare the WEE approach to current practice through a real dataset of patient fecal occult blood test outcomes at laboratories in Ontario as well as simulated data. The study approach improves precision of the estimates and the power of a hypothesis test to compare estimates in order to reduce the risk of classifying a laboratory as compliant or non-compliant in error.

Keywords: Statistical estimation • Hypothesis tests • Small samples • Laboratory compliance • FOBT • Proficiency tests

Abbreviations: FOBT: Fecal Occult Blood Test; LR: Likelihood Ratio, WEE: Weighted Estimating Equations

Introduction

In the United States, the Centers for Medicare and Medicaid Services regulate all laboratories testing performed on humans through the Clinical Laboratory Improvement Amendments. In Ontario, Canada, the Institute for Quality Management in Healthcare is an independent agency with a provincial mandate to assess the ability of laboratories to perform medical testing. To equip medical professionals with quality data for decisions impacting patient health, the mission of the regulatory agencies is to provide rigorous, objective, third-party evaluation of the medical diagnostic testing systems according to international standards. Various laboratories may be performing the same test; however, differences between test kits, measurement methods, and processes, for example, can contribute to measurable differences between the mean outcomes at the various laboratories. Non-compliance and unfavorable performance have important implications for licensing continuance and for attracting patients. Considering a test result that is either "positive" or "negative", a regulator or stakeholder may want to:

- Estimate the positive rate by laboratory
- Compare the positive rates across all laboratories
- Detect those laboratories which have a higher positive rate than their peers

The particular application under study relates to proficiency testing of laboratories which test samples for indications of colorectal cancer. In Ontario, the Colon Cancer Check program was initiated in 2008 as the first population-based, province-wide, organized screening program designed to raise screening rates and reduce deaths from colorectal cancer. Those individuals who are deemed to be at risk for developing colorectal cancer are encouraged to have a Fecal Occult Blood Test (FOBT) every two years. Studies show that when detected early by a FOBT, there is an estimated 90% likelihood of curing colorectal cancer [1]. A kit is provided to the patient who draws the FOBT sample at home and returns their sample to a laboratory for testing. At the laboratory, a technician tests the sample by a nominal examination system which is common to all laboratories. The nominal examination results in either a negative or positive test result. The FOBT test is normally negative. A positive result indicates that abnormal bleeding is occurring and informs the medical professional to order follow-up tests. Unlike most other diagnostic tests, oversight of the seven laboratories testing FOBT samples in Ontario is assigned to a committee ("FOBT committee") comprised exclusively of laboratory representatives.

This paper highlights shortcomings with the FOBT committee's approach to laboratory performance monitoring and suggests a more effective approach. The approach as of May 2014 ("Ontario FOBT proficiency test") is as follows. Monthly, each of the seven laboratories report their positive rate which is calculated as the number of samples having a positive test result relative to the total number of samples tested. The observed positive rate in the present month for each laboratory is compared to an acceptance interval and a rate outside this interval indicates that the laboratory is in potential non-compliance. Three consecutive months of this status prompts a letter of concern from the committee and can escalate to requests for re-training, peer visits, or a recommendation to the Ministry of Health that the non-compliant laboratory cease performing tests. The acceptance interval is determined by three standard deviations above and below the 12-month moving average

***Address for Correspondence:** Patricia Cooper Barfoot, University of Waterloo, 524 Chancery Lane , Waterloo, ON, Canada N2T 2N4, Tel: +226-338-3804; E-mail: triciabarfoot@gmail.com*

Received: 12 April, 2021, Manuscript No. jbmbs-21-29465; **Editor assigned:** 14 April, 2022, PreQC No. P-29465; **Reviewed:** 17 January, 202, QC No. Q-29465; **Revised**: 23 January, 2022, Manuscript No. R-29465; **Published:** 31 January, 2022, DOI: 10.37421/2155-6180.2022.13.90

of outcomes across all seven laboratories. As the observed positive rate for each laboratory is compared to the acceptance interval, no consideration is given to the uncertainty of the rate resulting from sample size (number of tests conducted). Further, the number of patients tested at the various laboratories may vary to a large extent. Depending on how the acceptance interval is calculated, the data observed at a larger laboratory could have larger influence on the acceptance interval than a small laboratory. Changes in performance at a larger laboratory could move the acceptance interval over time and a smaller laboratory that experiences no change may become non-compliant relative to the latest acceptance interval. Sample size at a particular laboratory has an important impact on the probability that the laboratory is classified as non-compliant relative to its peers.

A common approach to reduce the impact of sample size is to pool data across multiple time periods (for example, a year). While pooling data improves the precision of estimates and power of hypothesis tests, this approach increases bias and reduces power when the positive test rate changes over time. We expect that true positive rates by laboratory change slowly over time due to the effects of unobserved factors. Including historical data to reduce uncertainty due to small sample size trades bias for precision. Too much change in the parameter over time results in a large amount of bias and this trade-off is not beneficial. The decision whether to use present time data only or to include some or all observed historical data depends on the bias/variance trade-off.

There is an opportunity to improve the bias and uncertainty in estimates of positive rate beyond the approach of pooling data. The Weighted Estimating Equations (WEE) approach [2] borrows information from the past in order to manage a bias/variance trade-off in an estimate of present performance. Where stakeholders pay regular attention to laboratory performance issues, we expect that laboratory performance changes slowly over time. Some laboratory sample sizes may be small. The WEE approach increases the statistical information for estimation by involving all relevant past data through the weighted estimating equation. Using WEE, estimates of present performance have less bias than pooling data and less uncertainty than using present data only. The WEE has intuitive properties which can be understood by laboratory performance stakeholders.

We further note that the Ontario FOBT proficiency test approach to assess a particular laboratory's positive rate gives no consideration to patient-level risk factors. The underlying presumption is that all patients who present their samples for testing at the various laboratories have the same chance of a positive test outcome. Clearly, this assumption is questionable. For example, labs may serve different geographic regions where patient population test positivity rates are not directly comparable. While the dataset which motivates the work of this paper does not include covariates, note that the proposed approach may be easily extended to compare outcomes which are risk-adjusted for patient-level risk factors.

We consider a real dataset of outcomes from the seven laboratories that perform the FOBT test in Ontario. Based on this dataset, the committee responsible for FOBT test oversight estimates the positive rates for the present month by laboratory and compares these through hypothesis tests. This paper compares the estimates and hypothesis tests by the WEE approach with the present approach as well as another naïve approach that pools data across all time periods without weights. The objective of the WEE approach is to reduce uncertainty in the estimates when sample sizes at some laboratories are small and manage the added bias caused by possibly slowly changing positive rate over time. Further, we discuss the treatment of patient-level risk factors in order to improve the comparison of positive rate estimates across laboratories by the WEE approach.

Data

Consider real test result data from the seven laboratories conducting the same Fecal Occult Blood Test (FOBT) in Ontario over time. Test results are observed from 863,898 patients in Ontario over the 18-month period from January 2014 to June 2015. The patients are not necessarily the same over time and are not identified. The positive or negative test result is recorded by patient, by laboratory, and by month. Figure 1 gives the observed number of patients by time period and the observed positive rate over time across all laboratories.

Figure 1 shows that the positive rate across all laboratories drifts slowly over time. Changes to the positive rate may occur due to many complex factors; examples include a change in the distribution of samples across the laboratories, continuous improvement in the test process at one or more laboratories, and changes in test equipment at one or more laboratories. We do not want to assume a stochastic or deterministic model to describe the change in rate since it may be difficult to model the contributing factors and the model may only be useful for a short period of time. Instead, we want to estimate the rate assuming that it changes slowly over time in an unpredictable way.

A physician recommending a FOBT usually refers their patient to a particular laboratory for testing. In Ontario, a laboratory may serve patients from as few as 100 or as many as several thousand referring physicians. As such, the number of samples tested by time period varies considerably from laboratory to laboratory. Figure 2 gives the sample size and observed positive rate of each FOBT laboratory in the present time period (June 2015).

Figure 2 shows that there are large differences in the numbers of patients tested across the various laboratories. The number of FOBT samples tested varies from approximately 600 per time period to approximately 20,000 per time period. Due to this wide disparity in sample sizes, the power of the Ontario FOBT proficiency test to correctly classify small laboratories as compliant or non-compliant is a concern.

This paper proposes a way to improve the estimate of the mean positive rate at each of the laboratories, noting the wide disparity in sample sizes, so that comparisons of rates across the laboratories are more reliable. In Appendix A, we provide a set of mathematical notation for the observations of test results for each patient, at each laboratory, and at each time period. Further, we assume that the observations follow a binomial distribution which is frequently used to model the number of successes and number of failures in a sample. We describe the (7×1) vector of parameters which we need to estimate and we refer to this vector as θ. With an estimate for θ, which we call

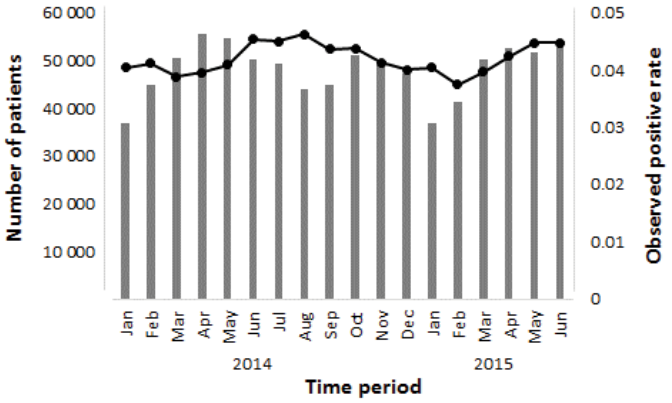

Figure 1. Sample size and observed positive rate (●) of FOBT laboratories in Ontario.

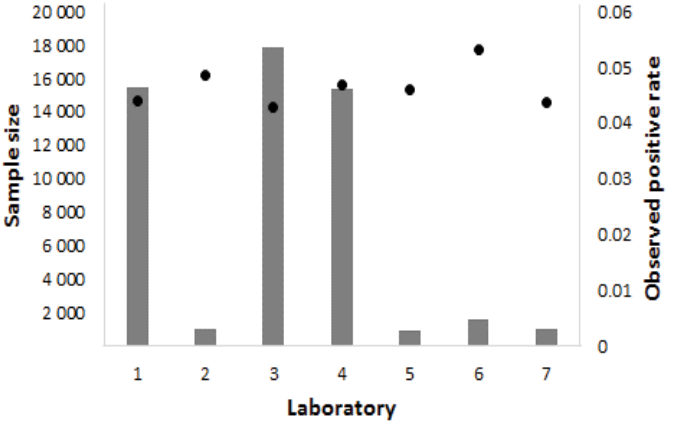

Figure 2. Sample size and observed positive rate (●) of FOBT laboratories in Ontario in June 2015.

$\hat{\theta}$, we can calculate estimates for the mean positive rate at each laboratory through (3) in Appendix A.

Methodology

There are two naïve approaches to estimate the parameter θ that are commonly used. One approach is to estimate θ using the observations from the latest time period only. In this paper, we refer to this as the 'naïve, present data only' approach. The drawback with this approach is that the present sample size may be small at some laboratories and so the estimate $\hat{\theta}$

has high variance. Another option to estimate θ is to use data across many time periods without regard for the time period of the data. In this paper, we refer to this as the 'naïve, all historical data' approach. Since more data are used for estimation, the variance of the estimate $\hat{\theta}$ is lower than when only present data are used. However, in the problem at hand where the positive rates at the various laboratories may be changing slowly over time, the estimate $\hat{\theta}$ may lead to biased estimates of the positive rates at various laboratories due to the influence of older data.

Weighted estimating equations approach

The Weighted Estimating Equations (WEE) approach offers a trade-off between estimation of positive rate using present time data only (naïve, present data only approach) or pooling the data by laboratory over all months (naïve, all historical data approach). This trade-off is especially important in the FOBT positive rate problem since sample sizes at some time periods may be small and the positive rate by laboratory may drift slowly over time in an unpredictable way. The premise behind the WEE approach is to estimate the present value of the parameter θ using all historical data, but to down-weight the influence of historical data. This is done by way of a function (called an 'estimating function') to calculate the estimate θ of parameter θ that involves all historical data as well as a set of weights. We use the notation t to denote the set of data observed for a time period t and w_t to denote the value of the weight which is selected for time period t, where t can take a value from 1 to T. Time period Tis the present (latest) time period. The estimating equation to calculate $\hat{\theta}$ is

$$w_1\psi_1\left(\hat{\theta};y_1\right)+w_2\psi_2\left(\hat{\theta};y_2\right)+\ldots+w_T\psi_T\left(\hat{\theta};y_T\right)=0 \qquad (1)$$

Where each $\psi_t(\hat{\theta};y_t)$is a score function which involves dataset y_t and $\hat{\theta}$ as given in Appendix B. To down-weight the influence of historical data, the various functions by time period are given weights which decline for time periods in the further past as $w_1 < w_2 < \cdots < w_T$Relative values of the weights control the trade-off between bias and variance of the estimate $\hat{\theta}$ and so the weights require careful selection. Since we expect that one or more the laboratory means drift slowly with time, we use weights that decrease exponentially for time periods further in the past. In particular, we propose a weight parameter which we refer to as λ and define the weight for each time period $t = 1, \ldots, T$ in terms of λ as

$$w_t = \lambda\left(1-\lambda\right)^{T-t} \qquad (2)$$

The weight parameter can take a value between 0 and 1 as $0 < \lambda < 1$. With (2), the weight for the most recent time period is proportional to λ, the time period before that has weight proportional to λ (1-λ), the time period before that λ (1-λ)^2 and so on. For convenience, we can divide each weight by the same constant $\sum_{t=1}^{T}\lambda(1-\lambda)^{T-t}$ so that$\sum_{t=1}^{T}w_t = 1$. Other definitions of decreasing weights are possible. A larger value of λ increases the relative weight of y_t in the estimate of θ. There is subjectivity in the selection of λ, but the value λ =0.1 is chosen for this application. Note that once the weight values are selected, a closed form solution for the WEE estimate $\hat{\theta}$ is possible using the equations given in Appendices A and B.

Naïve approaches

We consider the formulation of the WEE approach in relation to those of the Ontario FOBT proficiency test (naïve, present data only approach) and the approach that pools data across all time periods (naïve, all historical data approach). We point out that these two naïve approaches are special cases of the WEE approach with particular selections of the weights.

- Naïve, present data only approach: We can show that the WEE estimate simplifies to the usual calculation of a rate using only the latest time period data when we give all of the weight to the most recent time period (λ approaches 1 in (2) so that $w_T \approx 1$and $w_T \approx 0$ for t < T). The estimating function involves the present time dataset y_T only which is the same as the Ontario FOBT proficiency test estimate for each of the seven laboratories.
- Naïve, all historical data approach: We can show that the WEE estimate simplifies to the usual calculation of a rate using all historical data weighted equally when we select weights which are equal across all time periods (λ approaches 0 in (2) so that $w_t \approx 1/T$ for all t = 1,.....,T)

Further, we can show that at the two particular selections of the weights described previously, the estimate of the variance of $\hat{\theta}$ in the WEE formulation gives the usual estimates of variance [2].

Hypothesis tests

We compare the positive rates across laboratories for the present time period and detect laboratories which have a higher positive rate than their peers through tests of hypotheses. We consider the test with null hypothesis ('H_0') *vs*. the alternative hypothesis ('H_A') as follows:

H_0: All laboratories have same positive rate for the present time period

H_A: At least one of the laboratories has a different positive rate than the others

If H_0 is rejected, then there is statistically significant evidence that there are differences between test results across the laboratories. The committee can review estimates from each of the laboratories and carry out follow-up analysis to identify the nature of the differences across laboratories. Two characteristics of the hypothesis test are considered:

- **Size:** upper bound on the probability that the test is rejected for values of the parameter in the region where the null hypothesis is true
- **Power,** $\beta(\theta)$**:** The probability that the test is rejected at a particular value of the parameter, θ

For tests with a select value of size, we want the power of the test to be as large as possible on alternative values of the parameter θ. The power of a test is limited by the number of observations and so we look for an approach with the highest power of the test for H_0 *vs*. H_A given some relatively small sample sizes by laboratory. Since increasing sample size increases the power of a hypothesis test [3], one can improve the power of a test by pooling data across time periods. However, including data observed in time periods before a change occurs reduces the power of the test aimed at detecting the change. The decision whether to use present time data only or to include some or all observed historical data depends on the sample size of the laboratory experiencing the change and the size of the change which are both unknown. We compare the powers of the tests based on the WEE approach and the two naïve approaches.

In Appendix C, we give the calculation for a test statistic which we call the WEE Likelihood Ratio (LR) test statistic for testing H_0 *vs*. H_A (7). We refer to this test statistic as$\widetilde{S_W}$. Note that the WEE LR test statistic involves all historical data, but down-weights the influence of historical data through the weights. The weights by time period in the test statistic are also given by (2) with the value of the weight parameter λ selected previously. As is done in statistical hypothesis testing, we reject or don't reject H_0 in favour of H_A by comparing a test statistic to a critical value from the distribution of the test statistic. Appendix C gives an approximation for the distribution of the WEE LR test statistic. We use the approximation for the distribution of this statistic to determine if there is significant evidence to reject the null hypothesis. Note that at the two limiting values of weight parameter described previously, the WEE LR test statistic

and the approximation for its distribution give the usual results using the naïve, present data only or naïve, all historical data approaches.

Results

Ontario FOBT dataset

We compare the WEE estimates and the two naïve estimates for the mean positive rate at each laboratory for the FOBT dataset. Figure 3 gives the estimates based on the WEE approach and the two naïve approaches. Figure 3 shows that estimate by the WEE approach have less uncertainty than estimates using present data only across all laboratories. The uncertainties of the WEE estimates are comparable to those of the naïve, all historical data estimates. The WEE estimates of positive rates agree closely to those of the naïve, all historical data approach and are substantially different than those of the naïve, present data only approach for laboratories 1, 2, 3, and 4. The closeness of the WEE and naïve, all historical data estimates indicates that there was little change in actual positive rates at these laboratories over the 18 time periods. The naïve, all historical data estimates and WEE estimates will have a bigger difference when there is some change in the positive rates across time.

We calculate the WEE LR test statistic $\widetilde{S_W}$ to test the null hypothesis of no difference across mean positive rates across laboratories *vs*. the alternative hypothesis for the FOBT dataset using the procedure described above. Details of these calculations are given in Supplementary Table 1. We compare the value of the test statistic $\widetilde{S_W} = 53.8$ to the critical value of an approximate distribution of this quantity from the x^2 distribution, $x^2_{0.05}(6) = 12.6$. Since $\widetilde{S_W}$ is larger than the critical value of its approximate distribution, there is statistically significant evidence to reject the null hypothesis in favor of the alternative hypothesis for a size 0.05 test. The reject/don't reject decision is the same for the hypothesis test by the naïve, all historical data approach; however, the naïve, present data only approach gives no evidence to reject the null hypothesis *vs*. the alternative hypothesis. We see that an approach based on the present time period data is less sensitive at detecting differences among laboratories for this dataset. This is the current industry practice (as of May 2014) among the committee that oversees FOBT laboratories in Ontario.

Since we reject the null hypothesis that all laboratories have the same positive rate at the present time, we refer to the estimates in Figure 3 to understand the nature of the differences across laboratories. The estimates show that the WEE estimate of positive rate at laboratory 6 is significantly higher than the rates at each of the other laboratories. Note that this difference is not significant through comparison of naïve, present data only estimates.

In addition to comparing the WEE LR test statistic to a critical value, it is useful to track the trend of $\widetilde{S_W}$, the statistic for the test of H_0 *vs*. H_A over time. The WEE LR test statistic at a particular time period is calculated with all observations up to the end of that time period and the same weight values used previously. The WEE LR test statistics for the Ontario FOBT dataset across the 18 time periods are given in Figure 4. Whenever the test statistic is higher than the critical value for the size 0.05 test, there is statistically significant evidence to reject the null hypothesis H_0 in favor of alternative H_A.

Figure 4 shows evidence to reject null hypothesis H_0 in favour of alternative H_A from April 2014 to June 2015 for a size 0.05 test. There are significant differences in the positive rate at one or more of the laboratories over this period. Note that this graph does not point to a particular laboratory and so there may be different outlier laboratory(s) from period to period. The graph points to a change at one of the laboratories that began around March 2014. Further, there is a downward trend that starts around September 2014. The downward trend from September 2014 to June 2015 suggests that positive rates across the laboratories are becoming more consistent with one another. A distinctive trend in the WEE LR test statistic should be investigated with the follow-up analysis discussed previously.

Simulation study

We simulate data that resembles the fecal occult blood test in Ontario dataset to study the power and biasedness of the size 0.05 tests of hypotheses by the various approaches. We compare the WEE approach to the two naïve approaches. We discuss the limitations of the results and the impact of certain characteristics of the data.

We simulate datasets with sample sizes similar to the Ontario FOBT laboratory problem where the number of samples per month ranges from 600 to 35,000 across the seven laboratories and the total sample size is 60,000 observations per month. Figure 5 gives the sample sizes for each laboratory which are the same for each of the 18 time periods (months).

Each simulated dataset contains observations by laboratory per time period over a period of 18 time periods. As stated, the objective of the analysis is to regularly assess a laboratory's ability to provide an acceptable standard of service by comparison with peers. Parallel changes at all laboratories simultaneously may be of interest, but are not addressed here. Each dataset is designed with positive rate at the first time period equal to 0.042 (4.2%) for each of the seven laboratories. Following the first time period, a change is introduced into a single laboratory and positive rates at the remaining laboratories are unchanged. We simulate a change at either the largest laboratory or the smallest laboratory in order to study the power and

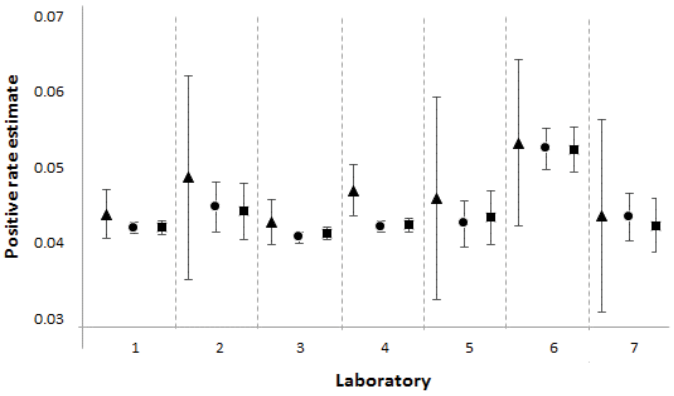

Figure 3. Estimates of positive rate for FOBT laboratories in June 2015 by naïve, present data only method (▲), naïve all historical data method (●), and WEE method (■). Vertical bars are the corresponding 95% upper control limits and lower control limits assuming normality.

Table 1. Simulation design profiles. The profile letters refer to the profiles for the positive rates over time period given in Figure 6.

Design Profile	Size of the Laboratory Undergoing the Change	Type of Change	Direction of Change	Positive Rate Profile (see Figure 6) for Laboratory m
I	Positive rate stay constant			Profile a for all m
II	Small laboratory	Step change	Increase	Profile a for m={1,...,6} Profile b for m=7
III	Small laboratory	Step change	Decrease	Profile a for m={1,...,6} Profile c for m=7
IV	Small laboratory	Linear change	Increase	Profile a for m={1,...,6} Profile d for m=7
V	Small laboratory	Linear change	Decrease	Profile a for m={1,...,6} Profile e for m=7
VI	Large Laboratory	Step change	Increase	Profile a for m={1,2,3,5,6,7} Profile b for m=4
VII	Large Laboratory	Step change	Decrease	Profile a for m={1,2,3,5,6,7} Profile c for m=4
VIII	Large Laboratory	Linear change	Increase	Profile a for m={1,2,3,5,6,7} Profile d for m=4
IX	Large Laboratory	Linear change	Decrease	Profile a for m={1,2,3,5,6,7} Profile e for m=4

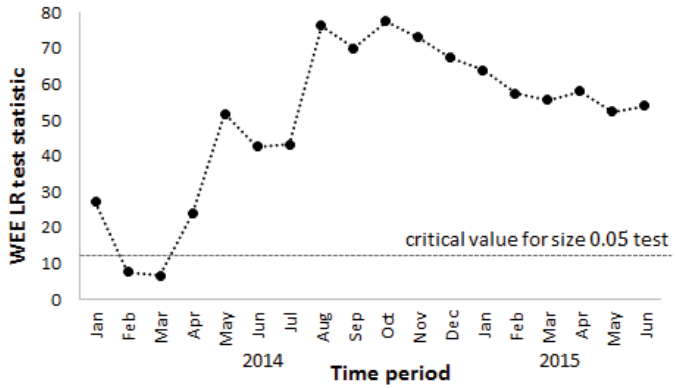

Figure 4. WEE LR test statistic ($\widetilde{S_W}$) to test H_0 *vs.* H_A by time.

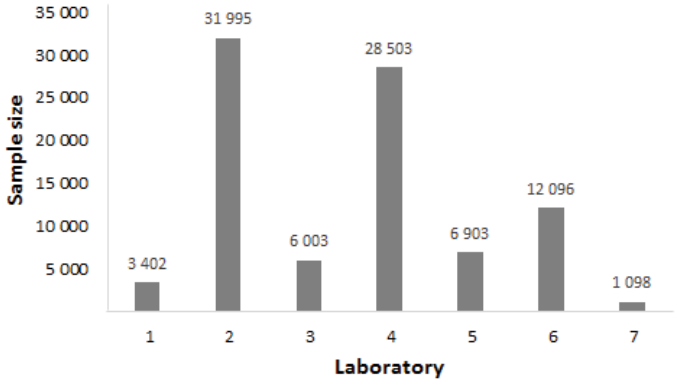

Figure 5. Sample size by laboratory per time period.

biasedness of the hypothesis test at the extremes of laboratory sample sizes. Many changes are possible; we simulate a step or linear change that increases or decreases the positive rate over the 18 time periods. We add a profile for the base case (called profile 'a') where the positive rate stays constant at all laboratories over time and a profile for each increasing/decreasing and step/linear positive rate change (called profiles 'b' through 'e'). Under these conditions, there are nine design profiles involving the positive rate changes for a large or small laboratory (called 'I' through 'IX') as summarized in Table 1. As noted, the profile letters in Table 1 refer to the profiles for the positive rates over time period given in Figure 6.

A positive or negative test response is simulated for each sample at each laboratory at each time period by the binomial distribution with the appropriate positive rate design value. The sample size for each laboratory by time period is given in Figure 5 and the positive rate design value by time period is given in Table 1 and Figure 6. We simulate 5000 datasets under these conditions for each of the nine design profiles.

For each of the 5000 simulated datasets for each of the nine design profiles, we calculate the WEE LR test statistic and reject or do not reject H_0 *vs.* H_A based on the approximation for its distribution under the null hypothesis as described in Appendix C. We do this for every value of the ending time period between 1 and 18 and for each of the naïve and WEE approaches to study the power and biasedness of the tests statistics by the various approaches over successive time periods. With these simulation results, we evaluate the size for the design profile where the null hypothesis is known to be true (profile I) and the power and biasedness for design profiles where the null hypothesis is known to be false (profiles II through IX). Figure 7 gives the percentage of test statistics by the WEE and naïve approaches where the null hypothesis is rejected at size 0.05 for data simulated by profile I where the null hypothesis is known to be true. Figure 7 shows that the percentage of tests rejected over time is similar for the WEE and naïve approaches. The WEE LR test statistic rejects the null hypothesis for 4.8% of datasets, and the naïve, all historical data and naïve, present data only approaches reject for 4.7% and 5.0% of datasets, respectively. The closeness of the observed sizes of the tests compared to the design value for the size of test (0.05 or 5%) is expected and indicates that the approximations for the critical values of the test statistics are reasonable. The observed differences in actual sizes of the tests among the three approaches do not have an important impact on the interpretation of the power of the tests to follow.

Figures 8-15 gives the percentage of LR test statistics by the WEE and naïve approaches where the null hypothesis H_0 is rejected in favour of the alternative H_A based on data simulated with each of the eight design profiles where the null hypothesis is known to be false. The graphs are interpreted as the observed power of the various test statistics to reject the null hypothesis with sizes of the test close to 0.05. We expect the observed power to increase

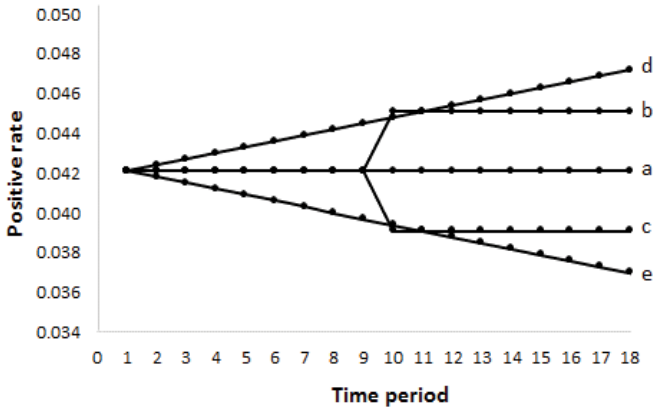

Figure 6. Positive rate design profiles.

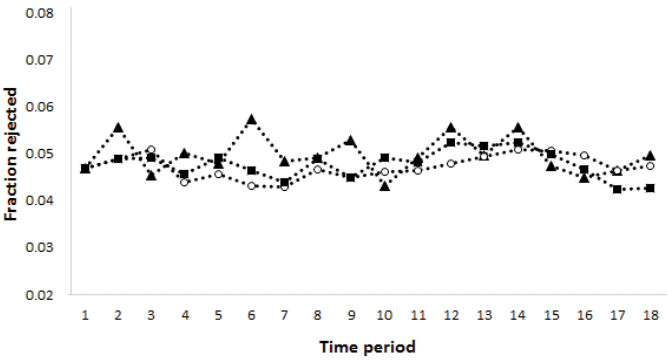

Figure 7. Percentage of tests of H_0 rejected for profile I (no change) by naïve method using present data only (▲), naïve method using all historical data (o), and WEE method (■).

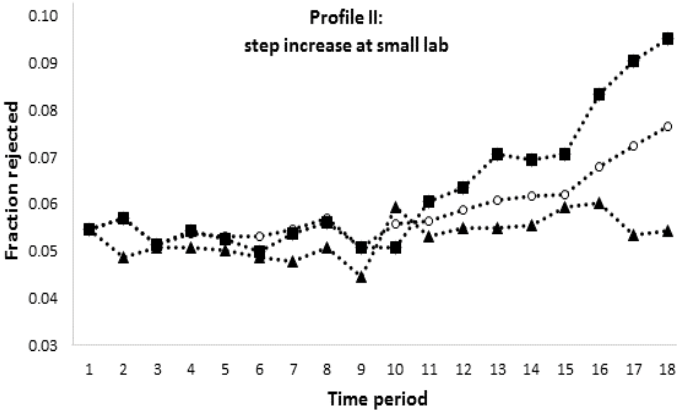

Figure 8. Percentage of tests of H_0 rejected for profiles II by naïve method using present data only (▲), naïve method using all historical data (o), and WEE method (■).

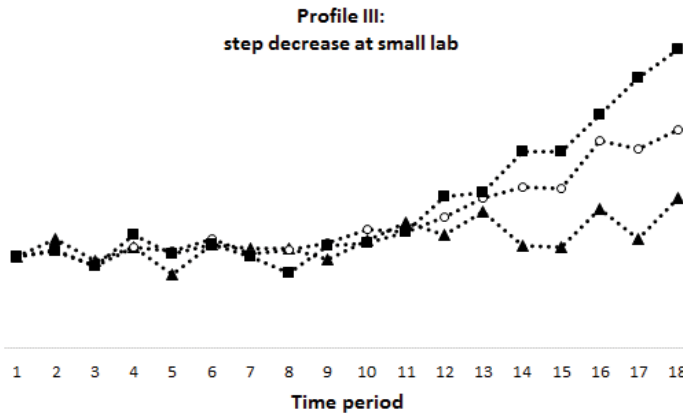

Figure 9. Percentage of tests of H_0 rejected for profiles III by naïve method using present data only (▲), naïve method using all historical data (o), and WEE method (■).

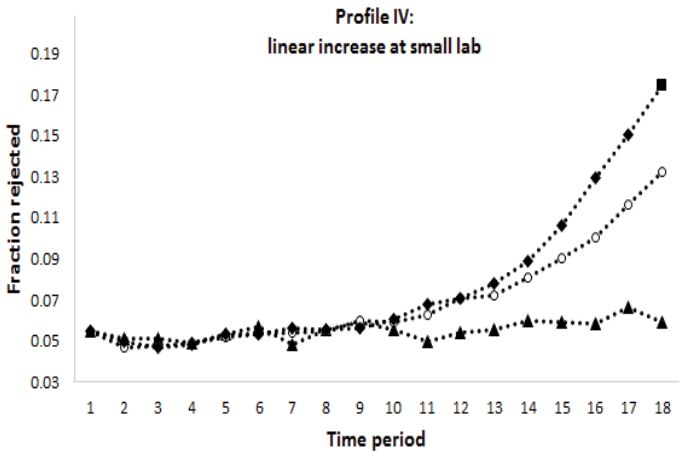

Figure 10. Percentage of tests of H_0 rejected for profiles IV by naïve method using present data only (▲), naïve method using all historical data (o), and WEE method (■).

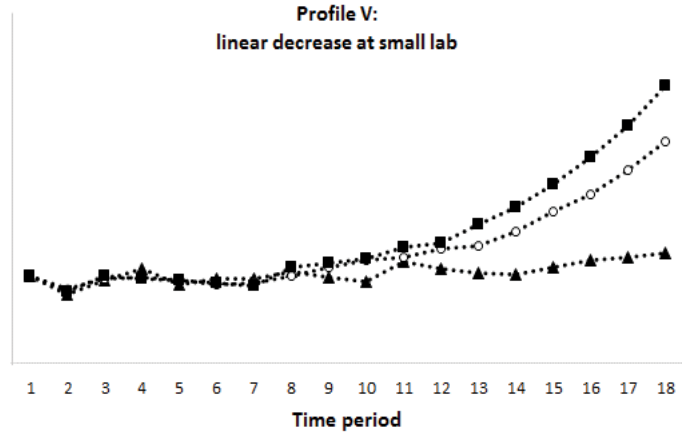

Figure 11. Percentage of tests of H_0rejected for profiles V by naïve method using present data only (▲), naïve method using all historical data (o), and WEE method (■).

according to the known change in positive rate. Note that there are differences in the scales of the vertical axes. Figure 8-15 shows that the power usually increases with more time periods since the step change and as the linear change gets larger. The exception is under the naïve, present data only approach where power does not increase with more time periods following a step change (naïve, present data only for profiles II, III, VI, VII). In general, the WEE approach has higher power to detect a change after a given number of time periods and requires fewer time periods to achieve a particular level of power.

We investigate how the power to detect a change increases as the size of the change increases in a follow-up simulation study. In this study, we simulate data where the true value of the positive rate does not change up to time period 9 and then either a linear or step change of various sizes occurs at the small laboratory 7. Figures 16 gives the observed power to detect the linear or step change of various sizes at three time periods following the change (that is, at time period 12) by test statistics from the various approaches.

Figure 16 shows that the power of the WEE approach to detect a step change of 0.024 in the positive rate (from 0.042 to 0.066) at laboratory 7 at three time periods since the change is favorable at 73% (0.73). As the size of the step change increases over the range from 0.003 to 0.024, the WEE approach has increasingly more power to detect the change than either naïve approach. Figure 17 shows that the power of the WEE approach to detect a linear change of 0.01 per time period (from 0.042 to 0.072) at laboratory 7 after

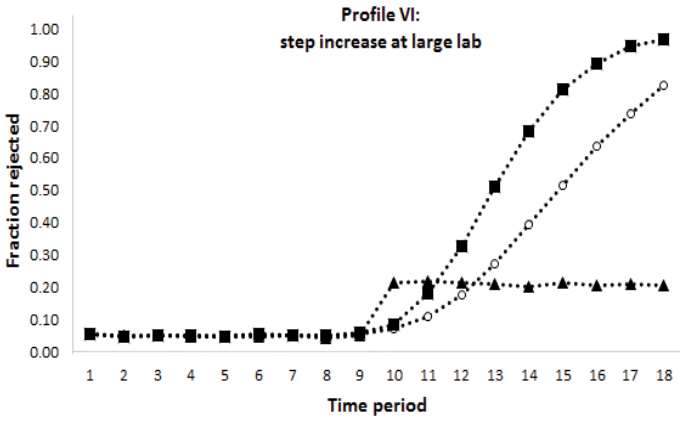

Figure 12. Percentage of tests of H_0 rejected for profiles VI by naïve method using present data only (▲), naïve method using all historical data (o), and WEE method (■).

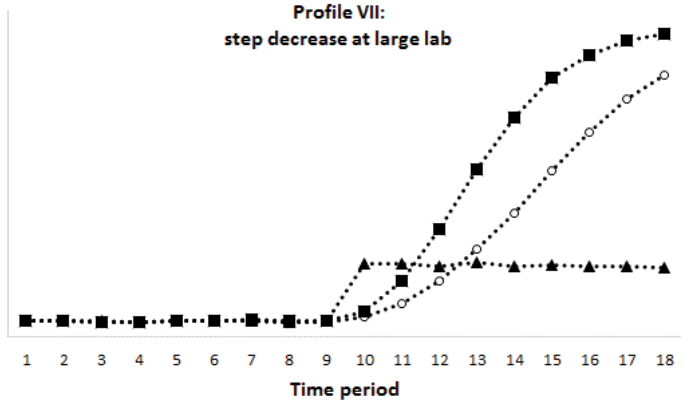

Figure 13. Percentage of tests of H_0 rejected for profiles VII by naïve method using present data only (▲), naïve method using all historical data (o), and WEE method (■).

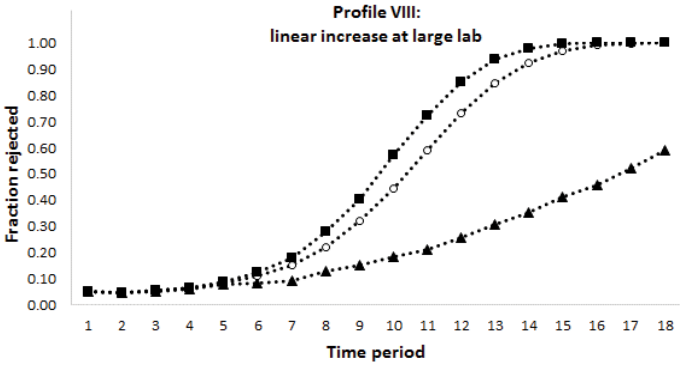

Figure 14. Percentage of tests of H_0 rejected for profiles VIII by naïve method using present data only (▲), naïve method using all historical data (o), and WEE method (■).

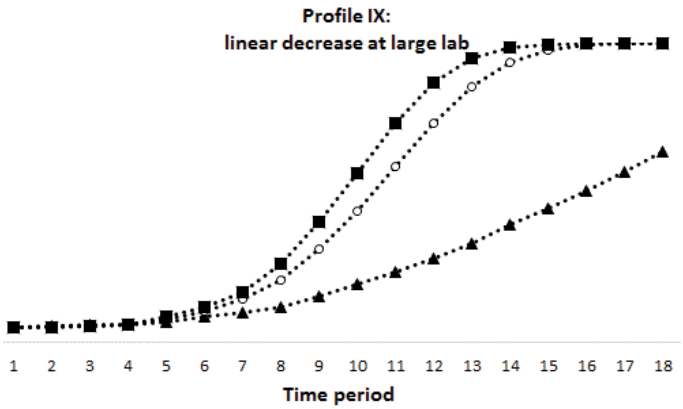

Figure 15. Percentage of tests of H_0 rejected for profiles IX by naïve method using present data only (/), naïve method using all historical data (o), and WEE method (■).

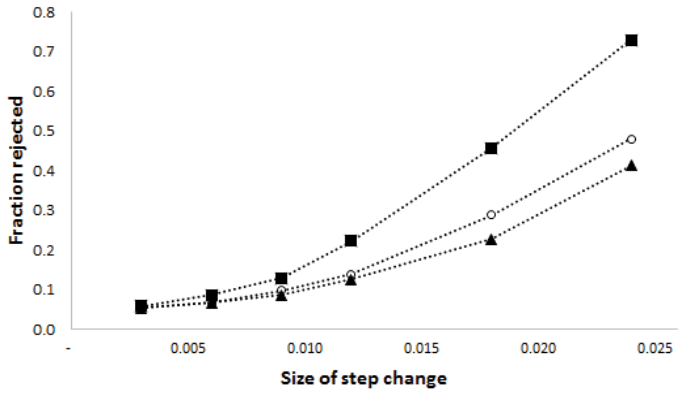

Figure 16. Power of test to detect change at a small laboratory after three time periods; following a step change.

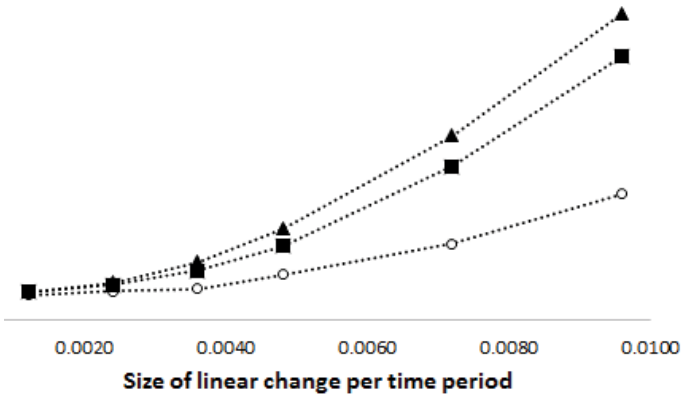

Figure 17. Power of test to detect change at a small laboratory after three time periods; following a linear change by naïve method using present data only (o), naïve method using all historical data (▲), and WEE method (■).

three time periods is 55% (0.55). The naïve present data only approach has slightly more power than the WEE approach for this example since the change is relatively large and the total amount of data since the change is relatively small compared to that from before the change. As more time passes since the start of the change, we expect that the power of the WEE approach to detect a change will surpass the power of the naïve; present data only approach, but exploring this through simulation remains as future work. The nine time periods of data observed before the change considerably reduce the power of the naïve approach using all historical data weighted equally compared to the other approaches. This study shows that there is favorable power of the WEE approach for detecting a change at a small laboratory within a short time frame depending on the size of the change.

Summary and Discussion

The proficiency test to assess the ability of laboratories to perform Fecal Occult Blood Tests (FOBT) in Ontario compares the observed positive rate from various laboratories to an acceptance interval based on data across all laboratories. When one or more of the laboratories test few samples relative to the other laboratories, the probability that the Ontario FOBT proficiency test incorrectly classifies the laboratory acceptable or non-compliant may be significant. There is wide disparity in sample sizes between the seven laboratories testing FOBT in Ontario. The power of the Ontario FOBT proficient test to correctly classify small laboratories is a concern.

We analyze real Ontario FOBT outcome data from seven laboratories over a period of 18 months. We use the Weighted Estimating Equations (WEE) approach since we expect that test performance may drift slowly over time in an unpredictable way and some sample sizes are small. We compare the WEE estimate of the positive rate to estimates by naïve approaches based on present time data only and all historical data weighted equally. The various approaches produce positive rate estimates that vary considerably from one another. The WEE estimate has similar precision to the naïve, all historical data estimate. Based on the WEE and the naïve, all historical data approaches, we reject a test of the null hypothesis that all laboratories have the same positive rate in favor of an alternative hypothesis that not all laboratories have the same positive rate. We do not reject this null hypothesis based on the analysis of present data only. Similarly, two of the tests against laboratory-specific alternative hypotheses are rejected based on the WEE approach and the naïve, all historical data approach, but not rejected based on the naïve, present data only approach. There are important differences in the results based on the WEE approach and those based on current industry practice.

We explore the power of the hypothesis test to detect difference between laboratories by the various approaches through simulated data designed with varying changes in positive rate over time. The conditions for the simulation study reflect those of the Ontario FOBT proficiency test at May 2014 as a prototype example, including the number of laboratories, sample sizes by laboratory, and initial positive rates. In general, the WEE approach has higher power to detect a change after a given number of time periods and requires fewer time periods to achieve a particular level of power. As the size of a step change increases, the WEE approach has increasingly more power to detect the change than either naïve approach under the particular simulation conditions. Under a linear change, initially the naïve, present data only approach has higher power but the power of the WEE approach surpasses its power within a short time frame depending on the size of the change.

Formal process monitoring could be used to provide quicker detection of small sustained shifts in positive rate and control the rate that laboratories are classified as non-complaint in error. A control chart statistic based on the likelihood ratio test can be used to detect a change in both the overall process mean and changes in the individual stream means [4]. The methodology to monitor multiple stream processes can be applied to monitor many laboratories simultaneously. The authors show that this test does not require a phase 1 sample where we collect data from an in-control process to set an appropriate control limit which saves cost. This work could be extended to develop a control chart for the WEE LR test statistic to improve time to detection of a small change and the probability of non-compliance misclassification [5-7].

Conclusion

The work of this paper indicates that a more reliable Ontario FOBT proficiency test can be constructed based on the weighted estimating approach that has suitable power to detect changes at a laboratory of any size and reduces the risk of classifying a laboratory as non-compliant in error.

Limitations and recommendations for future work

The work of this paper demonstrates the WEE approach based on the Ontario FOBT proficiency test which monitors anonymized patient results from all laboratories. It is important to point out two limitations. Firstly,

there are important laboratory factors that may contribute to differences across laboratories, such as differences in measurement methods, test kits, and laboratory processes. Further, important patient factors that may be contributing to differences, such as the foodstuffs consumed by the patient before taking the sample and the amount of time elapsed between drawing and testing a sample. We note that while data on these important laboratory factors and patient factors were not provided for the application at hand, risk-adjusted monitoring of outcomes through the WEE approach is possible. It should be noted however, that any factors relating to the quality of the laboratories should not be included whenever the results are intended to show differences in quality across laboratories. The importance of this work is to draw reliable comparisons across laboratory results. Differences that are detected can direct the appropriate follow-up work to understand the nature of the differences and drive quality improvement. Future work will extend this work to other datasets incorporating covariate data.

The second important limitation to note is that there are other methods of proficiency testing; for example, proficiency test which sends one or more artifacts between a number of participating laboratories. In this type of test, small samples may not be a concern. Consideration of present methods for estimating and comparing positive rates across laboratories through data collected by other types of proficiency tests remains as future work.

Acknowledgements

Financial support for this study was provided in part by research grant 105240 from the Natural Sciences and Engineering Research Council of Canada.

Conflict of Interest

The funding agreement disclosed above ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report. Each of the authors has no commercial associations that might create a conflict of interest in connection with this manuscript.

References

1. Cancer Care Ontario. "When caught early, colon cancer is more likely to be treated successfully." (2017).
2. Barfoot, Patricia L. Cooper, Stefan H. Steiner, and R. Jock MacKay. "Bias/Variance Trade-off in Estimates of a Process Parameter Based on Temporal Data." *J Qual Technol* 49 (2017):301-319. Google Scholar CrossRef Indexed At
3. Lehmann, Erich Leo and Joseph P. Romano. "Testing Statistical Hypotheses". NewYork: Springer Science+Business Media, USA, 3(2005).
4. Liu, Xuyuan, R. Jock MacKay, and Stefan H. Steiner. "Monitoring Multiple Stream Processes." *Qual Eng* 20 (2008):296-308. Google Scholar, CrossRef, Indexed at
5. Barfoot, Patricia Cooper, R. Jock MacKay, and Stefan H. Steiner. "Comparing and Monitoring Risk-adjusted Hospital Performance Measures: A Weighted Estimating Equations Approach." *MDM Policy & Practice* 3 (2008):1-12. Google Scholar CrossRef Indexed at
6. Small, Christopher G. "Expansions and Asymptotics for Statistics". Boca Raton: Chapman & Hall/CRC Press, USA, 2010.
7. Casella G., and R. L. Berger. "Statistical Inference, Second Edition". *Duxbury advanced se ries*, 2002.

How to cite this article: Barfoot, Patricia Cooper, Stefan H. Steiner and R. Jock MacKay. "Estimating and Comparing Diagnostic Laboratory Performance with Weighted Estimating Equations." J Biom Biostat 13 (2022): 90