Distribution Agreement

Signature:

Ganzhong Tian Date

Gestational Diabetes validation data study: an example of testing differences in sensitivity and specificity of two diagnostic methods using non-random sampling in a paired study design

By

Ganzhong Tian
Master of Science in Public Health

Biostatistics and Bioinformatics

Robert H. Lyles, PhD
(Thesis Advisor)

Glen A. Satten, PhD
(Thesis Reader)

Gestational Diabetes validation data study: an example of testing differences in sensitivity and specificity of two diagnostic methods using non-random sampling in a paired study design

By

Ganzhong Tian

M.Sc., Peking University 2013

B.Sc., Peking University 2010

Thesis Committee Chair: Robert H. Lyles, PhD

Reader: Glen A. Satten, PhD

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics
2017

Abstract

Gestational Diabetes validation data study: an example of testing differences in sensitivity and specificity of two diagnostic methods using non-random sampling in a paired study design

By Ganzhong Tian

Background: Comparing the accuracy of two diagnostic methods is a common problem in public health. This is especially the case when using a third 'gold standard' to determine patients' true disease status is either too expensive or time-consuming. For this kind of problem, it is highly desirable to have an efficient way of sampling and comparing the diagnostic properties of the two diagnostic methods.

Methods: In this study, we used a CDC study of Gestational Diabetes data as an example and considered an efficient design for validation sampling, which gives us more useful information regarding the diagnostic properties of two diagnostic methods for Gestational Diabetes. Also, to match with this sampling design, we proposed a new Wald test based on a 12-level multinomial distribution to compare the difference of the two diagnostic methods, in terms of some commonly evaluated diagnostic properties (e.g., sensitivity and specificity).

Computer simulation based on SAS/STAT and SAS/IML was used to implement the sampling process and assess the results of the hypothesis test, under different sampling designs and disease prevalence. We compared the results of the multinomial-based test against a more conventional McNemar's test, assumptions for which might be partially violated under our study setting and proposed sampling design.

Results: The results show that validation sampling only from the discordant pairs (those with disagreeing diagnostic results from the two diagnostic methods) will greatly boost the statistical power of testing the difference in sensitivity and specificity of the two diagnostic methods. Also, the Wald test we proposed performs well under different parameter settings and different sampling designs. In addition, our new test is superior to conventional McNemar's test in terms of statistical power and type-I error under the null hypothesis.

Gestational Diabetes validation data study: an example of testing differences in sensitivity and specificity of two diagnostic methods using non-random sampling in a paired study design

By

Ganzhong Tian

M.Sc., Peking University 2013
B.Sc., Peking University 2010

Thesis Committee Chair: Robert H. Lyles, PhD

Reader: Glen A. Satten, PhD

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics
2017

Acknowledgements

I would like to thank the Department of Biostatistics and Bioinformatics at Emory University for their support and guidance over the past two years. I would especially like to thank Dr. Robert Lyles for his willingness to be my thesis advisor and all the endless help and guidance that he gave me on my thesis, as well as his extensive knowledge and insight on this project. I would also like to thank Dr. Glen Satten not only for his idea and introduction of this thesis topic, but also for his willingness to be my thesis reader and all the suggestions he gave me on my thesis writing. I would also like to thank Dr. John Hanfelt for his valuable input and advice to this thesis project. Lastly, I would like to thank my family and friends for their love, support, and encouragement throughout this process.

Contents

List of Tables

List of Figures

Introduction

Diabetes that is newly diagnosed during pregnancy is called Gestational Diabetes or Gestational Diabetes Mellitus (GDM). This is a condition that often occurs around 24th week of pregnancy [**1**]. It can be further medically defined as any degree of glucose intolerance with onset or first recognition during pregnancy [**2**]. Also, the definition applies to those who underwent insulin or diet modification treatment for hyperglycemia during pregnancy, and it depends on whether the condition remains after pregnancy [**3**]. Like type-II diabetes, recent studies suggest that GDM happens when insulin resistance (IR) occurs during the pregnancy. This therefore leads to hyperglycemia since insulin's biological effect is to help the transfer of glucose from plasma into body cells, and insulin resistance prevents glucose from transferring into body cells [**4-6**]. Though GDM itself may be symptomless, pregnant women with the condition have increased risk of a series of pregnancy and delivery complications including pre-eclampsia, depression, hypertension and fetal macrosomia requiring a caesarean delivery. Newborns to mothers with poorly controlled gestational diabetes are at increased risk of being overweight and developing impaired glucose tolerance which may lead to type-II diabetes in the long term [**1-7**].

In the United States, the real prevalence of GDM has been poorly known until recently. Some articles suggest that the prevalence of GDM in the United States ranges from 1% to 14% of all pregnancies annually [**8-9**], while other studies on different sources of data estimated the real prevalence of GDM ranges from 1% to 25% [**10-12**]. Another authoritative study published by researchers from CDC in 2014 estimated the prevalence of GDM in United States to be 9.2% annually from 2007 to 2010 [**13**]. These publications showed the estimated prevalence of GDM depends on the source of data, and the diagnostic methods used for screening. Therefore, a method to assess and compare the accuracy of the data sources or diagnostic methods is desirable. In this study, the motivating example is to compare the diagnostic accuracy of Hospital Discharge data (HD) and Birth Certificate data (BC) as assessments of GDM status. Hospital discharge data are derived from a database system maintained and collected by the states. The states collect discharge records of all hospitalizations, which account for a very large fraction of births in the United States. The data collected include ICD-9 (or equivalent) codes, which give a standardized number for each condition or diagnosis. Hospitals usually compile these data for insurance and billing purposes, while the state gathers them as a surveillance system since they are useful to ascertain the health of the population [**14**].

In the United States, state laws require birth certificates to be completed for all births, and Federal law mandates national collection and publication of births and other vital statistics data. And each state has a registrar of vital records who oversees the upkeep of all vital records [**15-16**]. Based on these two sources of information, we can thus put patients into four separate 'cells' by looking at the diagnostic results of HD and BC. The four possible outcomes are: $BC^{+}\cap HD^{+}$(diagnosed positive both by HD and BC), $BC^{-}\cap HD^{+}$(diagnosed positive by HD but negative by BC), $BC^{+}\cap HD^{-}$ (diagnosed negative by HD but positive by BC), $BC^{-}\cap HD^{-}$(diagnosed negative both by HD and BC).

To investigate the accuracy of HD and BC, it is not enough to only know the diagnostic outcomes of the two data sources; we still need to know the true GDM status for at least some fraction of the patients. One often used 'gold standard' to determine whether a patient is truly GDM positive is called Medical Record Abstraction (MRA) [**17-18**]. Experienced medical abstractors are paid to go to the hospitals and look at a certain sample of the full hospital record of the women who were sampled into the study, to decide if the woman is truly GDM positive. Since we tell the abstractors which records to look at, we can specify which 'cells' we sample from. The 'gold standard' is very expensive and time-consuming to perform; thus, it is important to try to make the data obtained as informative as possible. One possible way to do this would be to control the sample size for patients from the four different 'cells', defined by BC and HD results, so that we can evaluate the accuracy of the two data sources using a relatively small sample size, or detect a smaller difference in accuracy. In particular, the cell $BC^{-}\cap HD^{-}$ is usually the largest cell but has the lowest prevalence of GDM. A random sample of pregnancies would include a large fraction of women from this potentially uninformative group.

The primary objective of this study is to assess the performance of an efficient sampling allocations (conditional on BC and HD status) and to assess the performance of a statistical test for comparing accuracy (sensitivity and specificity) of two diagnostic tests, when we have complete results of these tests, and can select which study subjects receive the ‘gold standard’ test. We are most interested in the situation in which the ‘gold standard’ is either too expensive or time-consuming to be performed on all subjects.

Methods

1. Data generation method

For proprietary reasons, the actual Birth Certificate (BC) as well as Hospital Discharge (HD) data from the motivating CDC-sponsored study are not currently available for our use. Thus, the comparison of the diagnostic properties of BC and HD as methods of assessing gestational diabetes will be discussed under different simulation scenarios using different hypothetical parameter settings.

Intuitively, regardless of the simulation approach and whether the data is real or simulated, we distinguish an overall sample size (N) from a validation sample size (n). All patients who are part of the study are contained in the overall sample and have BC and HD based assessments of GDM status. The overall sample BC and HD data can be displayed as a 2×2 contingency table as is shown in Table 1. Moreover, if we knew the true GDM status for all the study participants, we could then stratify the overall sample based on whether the patients are truly GDM positive. Then the 2×2 contingency table for the overall sample could be split into two separate 2×2 contingency tables, as are shown in Table 2. This would make it simple to estimate and compare the diagnostic properties (e.g., the sensitivities and specificities of the BC and HD assessments).

Considering a more realistic situation, because of limited resources, we simply cannot validate every patient in the overall sample. In this case, given a validation sample size n, where $n \leq N$, we are selecting n patients out of the overall sample of N as our sample for further statistical analysis. Then, again, it is very intuitive that the relationship of the overall sample and our sample can be displayed as in Table 3. Those sampled can be further stratified into two separate 2×2 contingency tables, as are shown in Table 4.

From Table 4, we can conclude any patient in the overall sample would eventually belong to one of the following twelve categories: (In A cell, but not sampled), (In A cell, sampled, with GDM+), (In A cell, sampled, with GDM-), (In B cell, but not sampled), (In B cell, sampled, with

GDM+), (In B cell, sampled, with GDM-), (In C cell, but not sampled), (In C cell, sampled, with GDM+), (In C cell, sampled, with GDM-), (In D cell, but not sampled), (In D cell, sampled, with GDM+), (In A cell, sampled, with GDM-). Therefore, as is shown in Table 4, each patient must belong to one of the twelve categories which can be listed as: a_1, a_0, a_u, b_1, b_0, b_u, c_1, c_0, c_u, d_1, d_0, d_u. Also, when we determine whether a patient should be sampled, we do not know her GDM status yet, so the event of sampling is independent of her GDM status. Hence, we have:

$$P(in\ a_1) = P(BC^+ \cap HD^+ \cap GDM^+ \cap sampled)$$
$$= P(sampled|BC^+ \cap HD^+)P(BC^+ \cap HD^+|GDM^+)P(GDM^+)$$
$$= P(sampled|BC^+ \cap HD^+)P(BC^+|HD^+ \cap GDM^+)P(HD^+|GDM^+)P(GDM^+) \quad (1)$$

$$P(in\ a_0) = P(BC^+ \cap HD^+ \cap GDM^- \cap sampled)$$
$$= P(sampled|BC^+ \cap HD^+)P(BC^+ \cap HD^+|GDM^-)P(GDM^-)$$
$$= P(sampled|BC^+ \cap HD^+)P(BC^+|HD^+ \cap GDM^-)P(HD^+|GDM^-)P(GDM^-) \quad (2)$$

$$P(in\ a_u) = P(BC^+ \cap HD^+ \cap not\ sampled)$$
$$= [1 - P(sampled|BC^+ \cap HD^+)]\{P(BC^+ \cap HD^+ \cap GDM^-) + P(BC^+ \cap HD^+ \cap GDM^+)\}$$
$$= [1 - P(sampled|BC^+ \cap HD^+)]\{P(BC^+|HD^+ \cap GDM^+)P(HD^+|GDM^+)P(GDM^+)$$
$$+ P(BC^+|HD^+ \cap GDM^-)P(HD^+|GDM^-)P(GDM^-)\} \quad (3)$$

$$P(in\ b_1) = P(BC^+ \cap HD^- \cap GDM^+ \cap sampled)$$
$$= P(sampled|BC^+ \cap HD^-)P(BC^+ \cap HD^-|GDM^+)P(GDM^+)$$
$$= P(sampled|BC^+ \cap HD^-)P(BC^+|HD^- \cap GDM^+)P(HD^-|GDM^+)P(GDM^+) \quad (4)$$

$P(in\ b_0) = P(BC^+ \cap HD^- \cap GDM^- \cap sampled)$

$= P(sampled|BC^+ \cap HD^-)P(BC^+ \cap HD^-|GDM^-)P(GDM^-)$

$= P(sampled|BC^+ \cap HD^-)P(BC^+|HD^- \cap GDM^-)P(HD^-|GDM^-)P(GDM^-)$ (5)

$P(in\ b_u) = P(BC^+ \cap HD^- \cap not\ sampled)$

$= [1 - P(sampled|BC^+ \cap HD^-)]\{P(BC^+ \cap HD^- \cap GDM^-) + P(BC^+ \cap HD^- \cap GDM^+)\}$

$= [1 - P(sampled|BC^+ \cap HD^-)]\{P(BC^+|HD^- \cap GDM^+)P(HD^-|GDM^+)P(GDM^+)$

$+ P(BC^+|HD^- \cap GDM^-)P(HD^-|GDM^-)P(GDM^-)\}$ (6)

$P(in\ c_1) = P(BC^- \cap HD^+ \cap GDM^+ \cap sampled)$

$= P(sampled|BC^- \cap HD^+)P(BC^- \cap HD^+|GDM^+)P(GDM^+)$

$= P(sampled|BC^- \cap HD^+)P(BC^-|HD^+ \cap GDM^+)P(HD^+|GDM^+)P(GDM^+)$ (7)

$P(in\ c_0) = P(BC^- \cap HD^+ \cap GDM^- \cap sampled)$

$= P(sampled|BC^- \cap HD^+)P(BC^- \cap HD^+|GDM^-)P(GDM^-)$

$= P(sampled|BC^- \cap HD^+)P(BC^-|HD^+ \cap GDM^-)P(HD^+|GDM^-)P(GDM^-)$ (8)

$P(in\ c_u) = P(BC^- \cap HD^+ \cap not\ sampled)$

$= [1 - P(sampled|BC^- \cap HD^+)]\{P(BC^- \cap HD^+ \cap GDM^-) + P(BC^- \cap HD^+ \cap GDM^+)\}$

$= [1 - P(sampled|BC^- \cap HD^+)]\{P(BC^-|HD^+ \cap GDM^+)P(HD^+|GDM^+)P(GDM^+)$

$+ P(BC^-|HD^+ \cap GDM^-)P(HD^+|GDM^-)P(GDM^-)\}$ (9)

$$P(in\ d_1) = P(BC^- \cap HD^- \cap GDM^+ \cap sampled)$$

$$= P(sampled|BC^- \cap HD^-)P(BC^- \cap HD^-|GDM^+)P(GDM^+)$$

$$= P(sampled|BC^- \cap HD^-)P(BC^-|HD^- \cap GDM^+)P(HD^-|GDM^+)P(GDM^+) \quad (10)$$

$$P(in\ d_0) = P(BC^- \cap HD^- \cap GDM^- \cap sampled)$$

$$= P(sampled|BC^- \cap HD^-)P(BC^- \cap HD^-|GDM^-)P(GDM^-)$$

$$= P(sampled|BC^- \cap HD^-)P(BC^-|HD^- \cap GDM^-)P(HD^-|GDM^-)P(GDM^-) \quad (11)$$

$$P(in\ d_u) = P(BC^- \cap HD^- \cap not\ sampled)$$

$$= [1 - P(sampled|BC^- \cap HD^-)]\{P(BC^- \cap HD^- \cap GDM^-) + P(BC^- \cap HD^- \cap GDM^+)\}$$

$$= [1 - P(sampled|BC^- \cap HD^-)]\{P(BC^-|HD^- \cap GDM^+)P(HD^-|GDM^+)P(GDM^+)$$

$$+ P(BC^-|HD^- \cap GDM^-)P(HD^-|GDM^-)P(GDM^-)\} \quad (12)$$

If we define the sampling rates as the probability of a patient being sampled for validation from the A, B, C, and D cells, and make them conditional on the cell status as φ_A, φ_B, φ_C, φ_D respectively, then we must have:

$$\varphi_A = P(sampled|in\ A\ cell) = P(sampled|BC^+ \cap HD^+)$$

$$\varphi_B = P(sampled|in\ B\ cell) = P(sampled|BC^+ \cap HD^-)$$

$$\varphi_C = P(sampled|in\ C\ cell) = P(sampled|BC^- \cap HD^+)$$

$$\varphi_D = P(sampled|in\ D\ cell) = P(sampled|BC^- \cap HD^-)$$

Also, we denote the diagnostic sensitivities and specificities of Birth Certificate data and Hospital Discharge data as:

$$SE_{BC} = P(BC^{+}|GDM^{+})$$

$$SP_{BC} = P(BC^{-}|GDM^{-})$$

$$SE_{HD} = P(HD^{+}|GDM^{+})$$

$$SP_{HD} = P(HD^{-}|GDM^{-})$$

Moreover, the sensitivities and specificities of Birth Certificate conditional on Hospital Discharge diagnostic status can be written as:

$$SE_{BC|HD^{+}} = P(BC^{+}|HD^{+}\cap GDM^{+})$$

$$SE_{BC|HD^{-}} = P(BC^{+}|HD^{-}\cap GDM^{+})$$

$$SP_{BC|HD^{+}} = P(BC^{-}|HD^{+}\cap GDM^{-})$$

$$SP_{BC|HD^{-}} = P(BC^{-}|HD^{-}\cap GDM^{-})$$

Then, we denote the prevalence of GDM as:

$$P(GDM^{+}) = \pi_{GDM}$$

Finally, we insert the above equations into equation (1) to equation (12), yielding equation (13) to equation (24):

$$P_{a_1} = P(in\ a_1) = P(sampled|BC^{+}\cap HD^{+})P(HD^{+}|GDM^{+})P(BC^{+}|HD^{+}\cap GDM^{+})P(GDM^{+})$$
$$= \varphi_A \times SE_{BC|HD^{+}} \times SE_{HD} \times \pi_{GDM} \qquad (13)$$

$$P_{a_0} = P(in\ a_0) = P(sampled|BC^{+}\cap HD^{+})P(HD^{+}|GDM^{-})P(BC^{+}|HD^{+}\cap GDM^{-})P(GDM^{-})$$
$$= \varphi_A \times \left(1 - SP_{BC|HD^{+}}\right) \times (1 - SP_{HD}) \times (1 - \pi_{GDM}) \qquad (14)$$

$$P_{a_u} = P(in\ a_u) = (1-\varphi_A)\times\left\{SE_{BC|HD^+}SE_{HD}\pi_{GDM} + \left(1-SP_{BC|HD^+}\right)(1-SP_{HD})(1-\pi_{GDM})\right\} \tag{15}$$

$$\begin{aligned} P_{b_1} = P(in\ b_1) &= P(sampled|BC^+\cap HD^-)P(HD^-|GDM^+)P(BC^+|HD^-\cap GDM^+)P(GDM^+) \\ &= \varphi_B\times SE_{BC|HD^-}\times(1-SE_{HD})\times\pi_{GDM} \end{aligned} \tag{16}$$

$$\begin{aligned} P_{b_0} = P(in\ b_0) &= P(sampled|BC^+\cap HD^-)P(HD^-|GDM^-)P(BC^+|HD^-\cap GDM^-)P(GDM^-) \\ &= \varphi_B\times\left(1-SP_{BC|HD^-}\right)\times SP_{HD}\times(1-\pi_{GDM}) \end{aligned} \tag{17}$$

$$P_{b_u} = P(in\ b_u) = (1-\varphi_B)\left\{SE_{BC|HD^-}(1-SE_{HD})\pi_{GDM} + \left(1-SP_{BC|HD^-}\right)SP_{HD}(1-\pi_{GDM})\right\} \tag{18}$$

$$\begin{aligned} P_{c_1} = P(in\ c_1) &= P(sampled|BC^-\cap HD^+)P(HD^+|GDM^+)P(BC^-|HD^+\cap GDM^+)P(GDM^+) \\ &= \varphi_C\times\left(1-SE_{BC|HD^+}\right)\times SE_{HD}\times\pi_{GDM} \end{aligned} \tag{19}$$

$$\begin{aligned} P_{c_0} = P(in\ c_0) &= P(sampled|BC^-\cap HD^+)P(HD^+|GDM^-)P(BC^-|HD^+\cap GDM^-)P(GDM^-) \\ &= \varphi_C\times SP_{BC|HD^+}\times(1-SP_{HD})\times(1-\pi_{GDM}) \end{aligned} \tag{20}$$

$$P_{c_u} = P(in\ c_u) = (1-\varphi_C)\times\left\{\left(1-SE_{BC|HD^+}\right)SE_{HD}\pi_{GDM} + SP_{BC|HD^+}(1-SP_{HD})(1-\pi_{GDM})\right\} \tag{21}$$

$$P_{d_1} = P(in\ d_1) = P(sampled|BC^{-}\cap HD^{-})P(HD^{-}|GDM^{+})P(BC^{-}|HD^{-}\cap GDM^{+})P(GDM^{+})$$

$$= \varphi_D \times \left(1 - SE_{BC|HD^{-}}\right) \times (1 - SE_{HD}) \times \pi_{GDM} \qquad (22)$$

$$P_{d_0} = P(in\ d_0) = P(sampled|BC^{-}\cap HD^{-})P(HD^{-}|GDM^{-})P(BC^{-}|HD^{-}\cap GDM^{-})P(GDM^{-})$$

$$= \varphi_D \times SP_{BC|HD^{-}} \times SP_{HD} \times (1 - \pi_{GDM}) \qquad (23)$$

$$P_{d_u} = P(in\ d_u) = (1 - \varphi_D) \times \left\{\left(1 - SE_{BC|HD^{-}}\right)(1 - SE_{HD})\pi_{GDM} + SP_{BC|HD^{-}}SP_{HD}(1 - \pi_{GDM})\right\}$$

$$(24)$$

Since each patient must inevitably fall into one of the 12 categories, we can further assume the random vector $(a_1, a_0, a_u, b_1, b_0, b_u, c_1, c_0, c_u, d_1, d_0, d_u)$ follows a 12-level multinomial distribution. Letting $\boldsymbol{Z} = (a_1, a_0, a_u, b_1, b_0, b_u, c_1, c_0, c_u, d_1, d_0, d_u)$, it is then apparent that:

$\boldsymbol{Z} \sim Multinomial(\left[P_{a_1}, P_{a_0}, P_{a_u}, P_{b_1}, P_{b_0}, P_{b_u}, P_{c_1}, P_{c_0}, P_{c_u}, P_{d_1}, P_{d_0}, P_{d_u}\right], N)$. So, the whole experiment of first sampling N patients and assessing their BC and HD status, then selecting n patients out of the overall sample of N and measuring the true GDM status of the sampled n patients can be simulated by using a random vector generator based on the 12-level multinomial distribution defined above.

2. Reparameterization and sampling schemes

Note that when we do the simulation based on a 12-level multinomial distribution, we need to set parameters as input, and based on the notations above, we need 11 unique parameters (overall sample size N is not included) to sufficiently describe the multinomial distribution, and per equation (13) to equation (24), these 11 input parameters are:

$\varphi_A, \varphi_B, \varphi_C, \varphi_D, SE_{BC|HD^+}, SE_{BC|HD^-}, SP_{BC|HD^+}, SP_{BC|HD^-}, SE_{HD},\ SP_{HD}, \pi_{GDM}$

Among them, $\varphi_A \sim \varphi_D$ are the sampling rates we plan to apply to patients from the A cell, B cell, C cell, and D cell defined in Table 1 respectively; π_{GDM} is the prevalence of GDM in the overall sample; and the sensitivities and specificities are either defined on Hospital Discharge or defined on Birth Certificate conditional on the diagnostic result of Hospital Discharge. But since our purpose of this study is to compare the sensitivity and specificity of Hospital Discharge against those of Birth Certificate, it is much better if we can re-parameterize the sensitivity and specificity parameters, in terms of $SE_{HD},\ SP_{HD}, SE_{BC},\ SP_{BC}$. To do this, let us first stratify the overall sample per the true GDM status, as is shown in Table 2. Pretending for the sake of argument that N_1 and N_0 represent the total GDM^+ and GDM^- target populations, by the definition of sensitivities we have:

$$SE_{BC} = P(BC^+|GDM^+) = P(BC^+ \cap HD^+|GDM^+) + P(BC^+ \cap HD^-|GDM^+) = \frac{A_1 + B_1}{N_1}$$

$$SE_{HD} = P(HD^+|GDM^+) = P(BC^+ \cap HD^+|GDM^+) + P(BC^- \cap HD^+|GDM^+) = \frac{A_1 + C_1}{N_1}$$

The odds ratio associating Hospital Discharge with Birth Certificate among those who are GDM^+ is defined in terms of conditional probabilities as below:

$$\psi_1 = \frac{P(BC^+ \cap HD^+|GDM^+)P(BC^- \cap HD^-|GDM^+)}{P(BC^+ \cap HD^-|GDM^+)P(BC^- \cap HD^+|GDM^+)} = \frac{A_1 D_1}{B_1 C_1}$$

Also, we have:

$$P(BC^+ \cap HD^+|GDM^+) + P(BC^+ \cap HD^-|GDM^+) + P(BC^- \cap HD^+|GDM^+) + P(BC^- \cap HD^-|GDM^+)$$
$$= \frac{A_1 + B_1 + C_1 + D_1}{N_1} = 1$$

To connect with our conditional sensitivities of Birth Certificate, we have:

$$SE_{BC|HD^-} = P(BC^+|HD^-\cap GDM^+) = \frac{P(BC^+\cap HD^-|GDM^+)}{P(BC^+\cap HD^-|GDM^+) + P(BC^-\cap HD^-|GDM^+)} = \frac{B_1}{B_1 + D_1}$$

And:

$$SE_{BC|HD^+} = P(BC^+|HD^+\cap GDM^+) = \frac{P(BC^+\cap HD^+|GDM^+)}{P(BC^+\cap HD^+|GDM^+) + P(BC^-\cap HD^+|GDM^+)} = \frac{A_1}{A_1 + C_1}$$

So, if we put together the above equations, we have:

$$SE_{BC|HD^-} + SE_{HD}\left(SE_{BC|HD^+} - SE_{BC|HD^-}\right) = \frac{B_1}{B_1 + D_1} + \frac{A_1 + C_1}{N_1}\left(\frac{A_1}{A_1 + C_1} - \frac{B_1}{B_1 + D_1}\right)$$

$$= \frac{A_1 + B_1}{N_1} = SE_{BC} \qquad (25)$$

And:

$$\frac{SE_{BC|HD^+}\left(1 - SE_{BC|HD^-}\right)}{\left(1 - SE_{BC|HD^+}\right)SE_{BC|HD^-}} = \frac{\frac{A_1}{A_1 + C_1}\left(1 - \frac{B_1}{B_1 + D_1}\right)}{\left(1 - \frac{A_1}{A_1 + C_1}\right)\frac{B_1}{B_1 + D_1}} = \frac{A_1 D_1}{B_1 C_1} = \psi_1 \qquad (26)$$

Using equation (25) and equation (26), we can use the quadratic formula to solve for the conditional sensitivities. The solutions are as follows:

$$SE_{BC|HD^-} = \frac{1 + (SE_{HD} - SE_{BC})(\psi_1 - 1) \pm \sqrt{4(1 - SE_{HD})SE_{BC}(\psi_1 - 1) + \left[1 + (SE_{HD} - SE_{BC})\left((\psi_1 - 1)\right)\right]^2}}{2(SE_{HD} - 1)(\psi_1 - 1)}$$

With $SE_{BC|HD^-} \in (0, 1)$.

And:

$$SE_{BC|HD^+} = \frac{1 + (SE_{HD} + SE_{BC})(\psi_1 - 1) \mp \sqrt{4(1 - SE_{HD})SE_{BC}(\psi_1 - 1) + \left[1 + (SE_{HD} - SE_{BC})\left((\psi_1 - 1)\right)\right]^2}}{2SE_{HD}(\psi_1 - 1)}$$

With $SE_{BC|HD^+} \in (0, 1)$.

Thus, given SE_{HD}, SE_{BC} and ψ_1, we can use these solutions to find $SE_{BC|HD^-}$ and $SE_{BC|HD^+}$.

Similarly, we can find two equivalent expressions for the specificities, by looking at the other 2×2 contingency table in Table 2, conditioning on GDM^-:

$$SP_{BC} = P(BC^-|GDM^-) = P(BC^-\cap HD^+|GDM^-) + P(BC^-\cap HD^-|GDM^-) = \frac{C_0 + D_0}{N_0}$$

$$SP_{HD} = P(HD^-|GDM^-) = P(BC^+\cap HD^-|GDM^-) + P(BC^-\cap HD^-|GDM^-) = \frac{B_0 + D_0}{N_0}$$

The odds ratio associating Hospital Discharge with Birth Certificate among those who are GDM^- is defined in terms of the conditional probabilities as:

$$\psi_0 = \frac{P(BC^+\cap HD^+|GDM^-)P(BC^-\cap HD^-|GDM^-)}{P(BC^+\cap HD^-|GDM^-)P(BC^-\cap HD^+|GDM^-)} = \frac{A_0D_0}{B_0C_0}$$

Similarly, we have:

$$P(BC^+\cap HD^+|GDM^-) + P(BC^+\cap HD^-|GDM^-) + P(BC^-\cap HD^+|GDM^-) + P(BC^-\cap HD^-|GDM^-)$$
$$= \frac{A_0 + B_0 + C_0 + D_0}{N_0} = 1$$

To connect with our conditional specificities of Birth Certificate, we have:

$$SP_{BC|HD^-} = P(BC^-|HD^-\cap GDM^-) = \frac{P(BC^-\cap HD^-|GDM^-)}{P(BC^+\cap HD^-|GDM^-) + P(BC^-\cap HD^-|GDM^-)} = \frac{D_0}{B_0 + D_0}$$

And:

$$SP_{BC|HD^+} = P(BC^-|HD^+\cap GDM^-) = \frac{P(BC^-\cap HD^+|GDM^-)}{P(BC^+\cap HD^+|GDM^-) + P(BC^-\cap HD^+|GDM^-)} = \frac{C_0}{A_0 + C_0}$$

So, if we put together the above equations, we have:

$$SP_{BC|HD^+} + SP_{HD}\left(SP_{BC|HD^-} - SP_{BC|HD^+}\right) = \frac{C_0}{A_0 + C_0} + \frac{B_0 + D_0}{N_0}\left(\frac{D_0}{B_0 + D_0} - \frac{C_0}{A_0 + C_0}\right)$$

$$= \frac{C_0 + D_0}{N_0} = SP_{BC} \qquad (27)$$

And:

$$\frac{SP_{BC|HD^-}\left(1 - SP_{BC|HD^+}\right)}{\left(1 - SP_{BC|HD^-}\right)SP_{BC|HD^+}} = \frac{\frac{D_0}{B_0 + D_0}\left(1 - \frac{C_0}{A_0 + C_0}\right)}{\left(1 - \frac{D_0}{B_0 + D_0}\right)\frac{C_0}{A_0 + C_0}} = \frac{A_0 D_0}{B_0 C_0} = \psi_0 \qquad (28)$$

Using equation (27) and equation (28), we can further have the solutions:

$$SP_{BC|HD^-} = \frac{1 + (SP_{HD} + SP_{BC})(\psi_0 - 1) \mp \sqrt{4(1 - SP_{HD})SP_{BC}(\psi_0 - 1) + [1 + (SP_{HD} - SP_{BC})(\psi_0 - 1)]^2}}{2SP_{HD}(\psi_0 - 1)}$$

With $SP_{BC|HD^-} \in (0, 1)$.

And:

$$SP_{BC|HD^+} = \frac{1 + (SP_{HD} - SP_{BC})(\psi_0 - 1) \pm \sqrt{4(1 - SP_{HD})SP_{BC}(\psi_0 - 1) + [1 + (SP_{HD} - SP_{BC})(\psi_0 - 1)]^2}}{2(SP_{HD} - 1)(\psi_0 - 1)}$$

With $SP_{BC|HD^+} \in (0, 1)$.

Thus, given SP_{HD}, SP_{BC} and ψ_0, we can use these solutions to find $SP_{BC|HD^-}$ and $SP_{BC|HD^+}$.

Based on the above results, the former 11 input parameters are reparametrized in terms of the following:

$\varphi_A, \varphi_B, \varphi_C, \varphi_D, \psi_0, \psi_1, SE_{BC}, SP_{BC}, SE_{HD},\ SP_{HD}, \pi_{GDM}$.

In this way, we can specify simulation scenarios in terms of the desired unconditional sensitivities and specificities, while generating multinomial samples based upon the probabilities defined in equations (13) through (24).

As was noted in the Introduction, the statistical power of the hypothesis tests for the difference in sensitivity and specificity between HD and BC may largely depend on the sampled discordant pairs. In other words, it should be the discordant pairs in the 2×2 contingency tables (e.g., Table 2) that contain useful information for the hypothesis testing. This observation suggests that a targeted sampling scheme focusing more validation effort in the B and C cells may be beneficial from the standpoint of statistical power.

To evaluate this assertion, in this study, we also wanted to examine the statistical power and behavior of the hypothesis tests under three major different sampling schemes. The three different sampling schemes are: Equally sampling from A, B, C, D cells (i.e., simple random sampling from the overall table); Equally sampling from A, B, C cells (i.e., simple random sampling from A, B, C cells only); and Equally sampling from B, C cells (i.e., simple random sampling from B, C cells only). The rationale here is to set the overall sample size as a fixed constant number, or as a random number but with a very small variance across all the simulations, under all these three sampling schemes. Thus, by switching from equally sampling from A, B, C, D cell to equally sampling from B and C cell only, with a fixed total sample size, we are increasing the numbers of discordant pairs. Because these contain the useful information for hypothesis testing, we thus can theoretically increase the power of the tests. By the same token, we should be able to detect a smaller difference of sensitivity or specificity with the same given resources. Alternatively, we could sample fewer patients for validation to achieve the same statistical goal.

3. Hypothesis testing for difference in sensitivity and specificity

To test the difference in sensitivity and specificity between the Hospital Discharge data and the Birth Certificate data, we used the data generation method described above and then applied statistical tests to the generated data, under different sampling schemes.

First, we considered the applicability of McNemar's test. As is shown in Table 4, eventually the simulated sample can be further stratified into two separate 2×2 contingency tables, by checking whether the true GDM status of the patients is positive. Then, we could apply a paired McNemar's test to the GDM^{+} stratum to test the difference in sensitivity between HD and BC; and we could then apply a similar paired McNemar's test to the GDM^{-} stratum to test the difference in specificity between HD and BC [19].

For the test to compare sensitivity, we define the following null and alternative hypotheses:

$$H_0\colon SE_{BC} = SE_{HD}\ vs.\ H_0\colon SE_{BC} \neq SE_{HD}$$

The test statistic for the difference in sensitivity (with continuity correction) was computed as [**19**]:

$$X^2 = \left(\left|\frac{b_1 - c_1}{2}\right| - \frac{1}{2}\right)^2 \Big/ \left(\frac{b_1 + c_1}{4}\right)$$

For a two-sided level-α test, if $X^2 > \chi^2_{1,1-\alpha}$ then we reject H_0; otherwise, if $X^2 \leq \chi^2_{1,1-\alpha}$ then we accept H_0.

For the test of specificity, we define:

$$H_0\colon SP_{BC} = SP_{HD}\ vs.\ H_0\colon SP_{BC} \neq SP_{HD}$$

The test statistic for the difference in specificity (with continuity correction) was computed as:

$$X^2 = \left(\left|\frac{b_0 - c_0}{2}\right| - \frac{1}{2}\right)^2 \Big/ \left(\frac{b_0 + c_0}{4}\right)$$

For a two-sided level-α test, if $X^2 > \chi^2_{1,1-\alpha}$ we reject H_0; otherwise, if $X^2 \leq \chi^2_{1,1-\alpha}$ we accept H_0.

A straightforward application of McNemar's test here would typically occur under the condition that the true conditions of the patients are known before applying the other two diagnostic tests to those with true positive condition (testing difference in sensitivity using McNemar’s test) or those with true negative condition (testing difference in specificity using McNemar’s test). In this sense, the McNemar’s test can be viewed as conditional on the true conditions of the patients. But in our study, since the results of the two diagnostic tests to be compared (BC vs HD) would be known already and the true GDM status would only be determined after the sampling, the conditions are somewhat different from those under which McNemar’s test might offer the best option. Thus, even though we had a paired result in the form of a 2×2 contingency table which could then be stratified by the true GDM status, the assumptions of McNemar’s test may be ‘partially violated’. Thus, McNemar’s test may lose power and/or validity relative to the more conventional pair-matched data case.

To clarify the above notion of ‘partially violated’, note that when we consider equal sampling from the cells (e.g., equally sampling from A, B, C, D cells, equally sampling from A, B, C cells, or equally sampling from B, C cells), the sampling rates in the B and C cells are always the same:

$$\varphi_B = \varphi_C$$

Take the test of the difference in sensitivity as an example. Per the equations of SE_{BC} and SE_{HD} we described in previous section and again pretending that N_1 and N_0 represent the entire target population, we have:

$$SE_{BC} = P(BC^+|GDM^+) = P(BC^+ \cap HD^+|GDM^+) + P(BC^+ \cap HD^-|GDM^+) = \frac{A_1 + B_1}{N_1}$$

$$SE_{HD} = P(HD^+|GDM^+) = P(BC^+ \cap HD^+|GDM^+) + P(BC^- \cap HD^+|GDM^+) = \frac{A_1 + C_1}{N_1}$$

Hence under the null hypothesis: $H_0: SE_{BC} = SE_{HD}$, the null hypothesis can be re-stated as:

$$H_0: B_1 = C_1$$

Then, per equation (16) and equation (19), we have:

$$P_{b_1} = \varphi_B \times SE_{BC|HD^-} \times (1 - SE_{HD}) \times \pi_{GDM} = \varphi_B \times \frac{B_1}{B_1 + D_1} \times \left(1 - \frac{A_1 + C_1}{N_1}\right) \times \pi_{GDM}$$

$$= \varphi_B \times \frac{B_1}{B_1 + D_1} \times \frac{B_1 + D_1}{N_1} \times \pi_{GDM} = \varphi_B \times \frac{B_1}{N_1} \times \pi_{GDM}$$

And:

$$P_{c_1} = \varphi_C \times (1 - SE_{BC|HD^+}) \times SE_{HD} \times \pi_{GDM} = \varphi_C \times \left(1 - \frac{A_1}{A_1 + C_1}\right) \times \left(\frac{A_1 + C_1}{N_1}\right) \times \pi_{GDM}$$

$$= \varphi_C \times \frac{C_1}{A_1 + C_1} \times \frac{A_1 + C_1}{N_1} \times \pi_{GDM} = \varphi_C \times \frac{C_1}{N_1} \times \pi_{GDM}$$

Then, it becomes apparent that when we are generating the data under the null hypothesis with $\varphi_B = \varphi_C$, we will always have $P_{b_1} = P_{c_1}$. This means that the chances to observe a patient falling into b_1 or c_1 is identical. In this case, if the data were generated under the null with $SE_{BC} = SE_{HD}$, the McNemar's test approach would still be valid, with an actual type-I error close to the chosen Alpha-level. The same conclusion holds for the McNemar's test of the difference in specificity.

However, note in contrast what happens when we are doing unbalanced sampling from the A, B, C, D cells, such that we are sampling from B and C cells with different sampling rates:

$$\varphi_B \neq \varphi_C$$

In this situation, even under the null hypothesis: $H_0: SE_{BC} = SE_{HD}$, we no longer have equal chances of observing a patient falling into b_1 or c_1. Thus, the McNemar's test statistic $X^2 = \left(\left|\frac{b_1 - c_1}{2}\right| - \frac{1}{2}\right)^2 / \left(\frac{b_1 + c_1}{4}\right)$ will have an expectation other than 0 ($E(X^2) \neq 0$), and the actual type-I error is either too small or too large compared with the chosen Alpha level. The same situation holds for the McNemar's test of the difference in specificity.

To fix this problem, we need to find another hypothesis testing method which can be used when $\varphi_B \neq \varphi_C$. So, we propose another test based on the 12-level multinomial distribution, and the rationale is developed as follows:

Per the equations mentioned above, the difference in sensitivity:

$$D_{SE} = SE_{HD} - SE_{BC} = \frac{A_1 + C_1}{N_1} - \frac{A_1 + B_1}{N_1} = \frac{C_1 - B_1}{N_1} = \frac{C_1}{N_1} - \frac{B_1}{N_1}$$

Because we can only observe the sampled data instead of the population, we do not know the true value of D_{SE}. But we can easily construct a natural estimator of $C_1 - B_1$, which is proportional to D_{SE}:

$$\widehat{D}_{SE} = \widehat{SE}_{HD} - \widehat{SE}_{BC} \propto C\frac{c_1}{c} - B\frac{b_1}{b} = (c_1 + c_0 + c_u)\frac{c_1}{c_1 + c_0} - (b_1 + b_0 + b_u)\frac{b_1}{b_1 + b_0}$$
$$= c_1\left(1 + \frac{c_u}{c_1 + c_0}\right) - b_1\left(1 + \frac{b_u}{b_1 + b_0}\right)$$

To obtain the estimator of its variance, we use the multivariate Delta method. The estimated form of the gradient vector ∇D_{SE} containing the 12 derivatives of $\widehat{D}_{SE}$ is shown as:

$$\widehat{\nabla D_{SE}}^T = \left(0, 0, 0, \frac{-b_u}{b_1 + b_0} + \frac{b_1 b_u}{(b_1 + b_0)^2} - 1, \frac{b_1 b_u}{(b_1 + b_0)^2}, \frac{-b_1}{b_1 + b_0}, \frac{c_u}{c_1 + c_0} - \frac{c_1 c_u}{(c_1 + c_0)^2} + 1, \frac{-c_1 c_u}{(c_1 + c_0)^2}, \frac{c_1}{c_1 + c_0}, 0, 0, 0\right)$$

We then have:

$$\widehat{Var}\left(\widehat{D}_{SE}\right) = \widehat{\nabla D_{SE}}\widehat{\boldsymbol{V}}\widehat{\nabla D_{SE}}^T ,$$

where $\widehat{\boldsymbol{V}}$ is the 12×12 variance-covariance matrix with multinomial structure, after replacing (P_{a_1}, P_{a_0}, P_{a_u},… P_{d_u}) by their natural estimates. These estimates are simply the number of observations in each cell divided by the overall sample size (N), for example, $\widehat{P}_{a_1} = a_1/N$:

$$\underset{\mathbf{12\times12}}{\widehat{\boldsymbol{V}}} = \begin{bmatrix} N\widehat{P}_{a_1}(1 - \widehat{P}_{a_1}) & -N\widehat{P}_{a_0}\widehat{P}_{a_1} & \cdots & -N\widehat{P}_{d_u}\widehat{P}_{a_1} \\ -N\widehat{P}_{a_1}\widehat{P}_{a_0} & N\widehat{P}_{a_0}(1 - \widehat{P}_{a_0}) & \cdots & -N\widehat{P}_{d_u}\widehat{P}_{a_0} \\ \vdots & \vdots & \ddots & \vdots \\ -N\widehat{P}_{a_1}\widehat{P}_{d_u} & -N\widehat{P}_{a_0}\widehat{P}_{d_u} & \cdots & N\widehat{P}_{d_u}(1 - \widehat{P}_{d_u}) \end{bmatrix}$$

Then a large sample-based Wald test based on this estimator of difference in sensitivity can be constructed as:

$$\mathrm{T}_{SE} = \frac{\left(\widehat{D}_{SE} - 0\right)^2}{\widehat{Var}\left(\widehat{D}_{SE}\right)} \sim \chi_1^2$$

Similarly, for the test of difference in specificity, another Wald test based on the estimator of a quantity that is proportional to the difference in specificity can be constructed as:

$$\mathrm{T}_{SP} = \frac{\left(\widehat{D}_{SP} - 0\right)^2}{\widehat{Var}\left(\widehat{D}_{SP}\right)} \sim \chi_1^2\,,$$

with the corresponding components, as follows:

$$\widehat{D}_{SP} = c_0\left(1 + \frac{c_u}{c_1 + c_0}\right) - b_0\left(1 + \frac{b_u}{b_1 + b_0}\right)$$

$$\widehat{Var}\left(\widehat{D}_{SP}\right) = \widehat{\nabla D_{SP}}\,\widehat{\boldsymbol{V}}\,\widehat{\nabla D_{SP}}^{T}$$

$$\widehat{\nabla D_{SP}}^{T} = \left(0, 0, 0, \frac{b_0 b_u}{(b_1 + b_0)^2}, \frac{-b_u}{b_1 + b_0} + \frac{b_0 b_u}{(b_1 + b_0)^2} - 1, \frac{-b_0}{b_1 + b_0}, \frac{-c_0 c_u}{(c_1 + c_0)^2}, \frac{c_u}{c_1 + c_0} - \frac{c_0 c_u}{(c_1 + c_0)^2} + 1, \frac{c_0}{c_1 + c_0}, 0, 0, 0\right)$$

We can conclude from the above equations, that no matter whether we are using McNemar's test or the Wald test based on the 12-level multinomial distribution, the test statistics only use information in the B cell and C cell. It supports our previous assertion that given the same total sample size, if we only sample from the B and C cells (i.e., we only sample those with disagreeing diagnostic results from HD and BC), we will have higher statistical power for hypothesis testing.

To further support our assertions, we first generated the data using SAS/STAT and SAS/IML [**20**] and computed the test statistics under the null, with huge number of replicates, to see if the probability of rejecting the null hypothesis is near the previously specified Alpha-level (type-I error). If the probability of rejecting the null for a given testing strategy is near the previously specified Alpha-level, the validity of that test is supported; otherwise, is not. The overall sample parameter settings for this experiment (parameters only describe the overall sample and do not affect the sampling) are listed as below:

$\pi_{GDM} = 0.10, SE_{HD} = 0.80, SE_{BC} = 0.80, SP_{HD} = 0.90, SP_{BC} = 0.90, \psi_0 = 2.111,\ \psi_1 = 2.667, N = 10000$

Given the overall sample parameters, for equally sampling from A, B, C, D cell, the sampling rates were set as:

$$\varphi_A = \varphi_B = \varphi_C = \varphi_D = 0.1500$$

Given the overall sample parameters, for equally sampling from A, B, C cell, the sampling rates were set as:

$$\varphi_A = \varphi_B = \varphi_C = 0.5719$$

Given the overall sample parameters, for equally sampling from B, C cell, the sampling rates were set as:

$$\varphi_B = \varphi_C = 0.8847$$

For all these sampling schemes, the total number of simulated replicates were set at 100,000 times. The selected sampling rates were chosen in order to make sure that the total sample sizes across different sampling schemes are very close (the expected sample size is 1500 in each case), so that results are more comparable across different schemes.

After the large sample exercise, we also wanted to investigate the performance of McNemar's test and our Wald test under different prevalence levels. To do this, we simulated the data under similar parameter settings:

$SE_{HD} = 0.80, SE_{BC} = 0.80, SP_{HD} = 0.90, SP_{BC} = 0.90, \psi_0 = 2.111,\ \psi_1 = 2.667, N = 10000$

But instead of keeping the SE_{HD} and SP_{BC} at a fixed level, in this series of simulations we increased SE_{HD}or decreased SP_{BC} by very small amounts each time, then performed 100,000 replicated simulations respectively under the varying conditions. For example, we first performed a simulation of 100,000 replicates under the null:

$SE_{HD} = 0.80, SE_{BC} = 0.80, SP_{HD} = 0.90, SP_{BC} = 0.90, \psi_0 = 2.111,\ \psi_1 = 2.667, N = 10000$

Then, after that, we increased SE_{HD} by 0.005, so the parameters were under the alternative hypothesis:

$SE_{HD} = 0.805, SE_{BC} = 0.80, SP_{HD} = 0.90, SP_{BC} = 0.90, \psi_0 = 2.111,\ \psi_1 = 2.667, N = 10000.$

And we then performed another simulation of 100,000 replicates under this condition. It allows us to see how the rejection rate (statistical power) of the hypothesis tests change and compare as the difference in sensitivity or specificity increases. Also, the whole simulation procedure was repeated under different levels of π_{GDM} so that we can also evaluate the impact of disease prevalence on the statistical tests' power.

Table 1. 2×2 contingency table of overall sample data.

		Hospital Discharge		
		Yes	No	
Birth Certificate	Yes	A	B	
	No	C	D	
				N

Where N stands for the size of the overall sample; A stands for number of patients diagnosed positive both in Hospital discharge data and Birth Certificate data; B stands for number of patients diagnosed positive in Birth Certificate data and diagnosed negative in Hospital discharge data; C stands for number of patients diagnosed positive in Hospital discharge data and diagnosed negative in Birth Certificate data; D stands for number of patients diagnosed negative both in Hospital discharge data Birth Certificate data. Thus, have: $A + B + C + D = N$.

Table 2. 2×2 contingency tables of overall sample data, stratified by GDM status.

GDM +		Hospital Discharge		
		Yes	No	
Birth Certificate	Yes	A_1	B_1	
	No	C_1	D_1	
				N_1

GDM -		Hospital Discharge		
		Yes	No	
Birth Certificate	Yes	A_0	B_0	
	No	C_0	D_0	
				N_0

Where $A_1 + B_1 + C_1 + D_1 = N_1$ and $A_0 + B_0 + C_0 + D_0 = N_0$

Table 3. 2×2 contingency tables of overall sample data, stratified by whether selected as validation sample.

Sampled=Yes		Hospital Discharge		
		Yes	No	
Birth Certificate	Yes	a	b	
	No	c	d	
				n

		Hospital Discharge		
Sampled=No		Yes	No	
Birth Certificate	Yes	A-a	B-b	
	No	C-c	D-d	
				N-n

Where $a + b + c + d = n$

Table 4. 2×2 contingency tables of overall sample data, stratified by GDM status and whether selected as validation sample.

		Hospital Discharge		
Sampled=No		Yes	No	
Birth Certificate	Yes	a_u=A-a	b_u=B-b	
	No	c_u=C-c	d_u=D-d	
				N-n

		Hospital Discharge		
Sampled=Yes, GDM+		Yes	No	
Birth Certificate	Yes	a_1	b_1	
	No	c_1	d_1	
				n_1

		Hospital Discharge		
Sampled=Yes, GDM-		Yes	No	
Birth Certificate	Yes	a_0	b_0	
	No	c_0	d_0	
				n_0

Where $a = a_0 + a_1$, $b = b_0 + b_1$, $c = c_0 + c_1$, $d = d_0 + d_1$, $n = n_0 + n_1$

Results

As is described in Methods, 100,000 simulations were first performed with respect to each different sampling scheme, while keeping the population parameters and expected number of sample size fixed. These results were used to check the type-I errors of the statistical tests.

Table 5 summarizes 100,000 simulations for equally sampling from the A, B, C, D cells, with an expected sample size at 1,500. From the type-I error we can see the McNemar's test has a smaller type-I error than 0.05. This can be explained by the relatively small number of b_1 and c_1 patients (the mean numbers of patients in b_1 and c_1 are less than 20), so the continuity corrected version of McNemar's test is not very satisfying. Instead, an exact binomial test would likely have better type-I error. Or, this situation may disappear if we have a much bigger overall sample size N or sample size n. Nevertheless, the Wald test maintains its type-I error near 0.05, which supports the validity of this statistical test.

Similar conclusions can be drawn from Table 6 and Table 7, which respectively summarize 100,000 simulations for equally sampling from the A, B, C cells and the B, C cells. The Wald test still performs well in both cases, and thanks to a much larger number of b_1 and c_1 patients observed, the continuity corrected version of McNemar's test is much better. This supports our assertion that the assumptions of the paired McNemar's test is 'partially violated', but that it can still be used in scenarios with equal B, C cells sampling. However, it still suggests our Wald test is superior in terms of type-I error.

As is shown in Figure 1 to Figure 4, we changed the difference in sensitivity from 0 to 0.150, by 0.005 each time, with prevalence of GDM equal to 0.05, 0.10, 0.20, 0.40, respectively, and then performed 100,000 simulations for each unique parameter setting under each one of the three equally sampling schemes. Because the simulations were performed under the alternative hypothesis, the rejection probabilities of the statistical tests are their statistical powers. From each of these 4 plots, we can easily conclude the following:

- First, our Wald test based on the 12-level multinomial distribution is always more powerful than the paired McNemar's test.
- Second, given the same overall and validation sample size, sampling only from B, C cells greatly improves the power of the test, compared with sampling from A, B, C cells or sampling from A, B, C, D cells.
- Third, as the prevalence of the disease increases, the power of the sensitivity tests also increases. And the lower the prevalence, the greater the benefit in power when sampling only from B/C cells.

Similarly, as is shown in Figure 5 to Figure 8, we changed the difference in specificity from 0 to 0.030, by 0.001 each time, with prevalence of GDM equals to 0.05, 0.10, 0.20, 0.40, respectively, and then performed 100,000 simulations for each unique parameter setting under each one of the three equally sampling schemes. The reason we chose to change the difference in specificity by 0.001 each time was because the prevalence of the disease is relatively low (less than 0.40), so the observed GDM condition negative stratum would be large, thus, the tests become more sensitive to the difference in specificity. Similarly, we can conclude the following based on these plots:

- First, our Wald test of difference in specificity based on 12-level multinomial distribution is always more powerful than the paired McNemar's test of the difference in specificity.
- Second, given the same sample sizes, sampling only from B/C cells improves the power of the test, compared with sampling from A/B/C cells or sampling from A/B/C/D cells.
- Third, as the prevalence of the disease increases, the power of the specificity tests decrease slightly. And the higher the prevalence, the greater the benefit in power when sampling only from B/C cells.

Table 5. Results of computer simulation comparing the type-I error of McNemar's test and Wald test, under the null hypothesis $H_0\colon SE_{BC} = SE_{HD}$ and $H_0\colon SP_{BC} = SP_{HD}$, with equal sampling from A, B, C, D cells.

Simulated result	Number of Simulations	Mean	Std Dev	Min	Max
A	100000	826.75	27.57	703	947
a_1	100000	100.33	9.95	63	148
a_0	100000	23.71	4.87	7	49
a	100000	124.04	11.04	79	173
B	100000	873.31	28.22	746	1001
b_1	100000	19.71	4.45	4	41
b_0	100000	111.27	10.49	69	157
b	100000	130.98	11.39	84	178
C	100000	873.29	28.18	748	1006
c_1	100000	19.70	4.44	5	41
c_0	100000	111.32	10.48	69	160
c	100000	131.02	11.37	84	185
D	100000	7426.65	43.66	7235	7630
d_1	100000	10.29	3.20	1	25
d_0	100000	1103.76	31.38	956	1237
d	100000	1114.05	31.51	966	1254
Sample size	100000	1500.09	35.70	1318	1644

Test Statistics	Number of Simulations	Mean	Std Dev	Min	Max	Type-I error
Test of difference in SE (McNemar)	100000	0.77	1.22	0	15.63	3.39%
Test of difference in SP (McNemar)	100000	0.90	1.34	0	17.72	4.23%
Test of difference in SE (Wald test)	100000	1.02	1.45	0	19.18	5.31%
Test of difference in SP (Wald test)	100000	1.00	1.43	0	18.83	5.04%

Table 6. Results of computer simulation comparing the type-I error of McNemar's test and Wald test, under the null hypothesis $H_0: SE_{BC} = SE_{HD}$ and $H_0: SP_{BC} = SP_{HD}$, with equal sampling from A, B, C cells.

Simulated result	Number of Simulations	Mean	Std Dev	Min	Max
A	100000	826.72	27.53	712	961
a_1	100000	382.57	19.18	297	463
a_0	100000	90.37	9.50	54	131
a	100000	472.94	21.21	387	564
B	100000	873.38	28.26	752	990
b_1	100000	75.07	8.66	40	119
b_0	100000	424.36	20.21	341	514
b	100000	499.43	21.86	414	595
C	100000	873.20	28.34	756	997
c_1	100000	75.07	8.62	42	116
c_0	100000	424.29	20.20	334	512
c	100000	499.35	21.81	410	591
D	100000	7426.70	43.78	7251	7618
d_1	100000	0.00	0.00	0	0
d_0	100000	0.00	0.00	0	0
d	100000	0.00	0.00	1E-10	1E-10
Sample size	100000	1471.73	35.56	1306	1621

Test Statistics	Number of Simulations	Mean	Std Dev	Min	Max	Type-I error
Test of difference in SE (McNemar)	100000	0.87	1.31	0	19.57	4.11%
Test of difference in SP (McNemar)	100000	0.95	1.38	0	18.82	4.68%
Test of difference in SE (Wald test)	100000	1.00	1.41	8E-11	20.49	4.94%
Test of difference in SP (Wald test)	100000	1.00	1.42	0	24.65	5.01%

Table 7. Results of computer simulation comparing the type-I error of McNemar's test and Wald test, under the null hypothesis $H_0: SE_{BC} = SE_{HD}$ and $H_0: SP_{BC} = SP_{HD}$, with equal sampling from B, C cells.

Simulated result	Number of Simulations	Mean	Std Dev	Min	Max
A	100000	826.63	27.53	709	953
a_1	100000	0.00	0.00	0	0
a_0	100000	0.00	0.00	0	0
a	100000	0.00	0.00	1E-10	1E-10
B	100000	873.28	28.16	742	1008
b_1	100000	116.15	10.68	69	164
b_0	100000	656.40	24.69	543	778
b	100000	772.55	26.65	657	913
C	100000	873.36	28.09	757	996
c_1	100000	116.11	10.72	72	160
c_0	100000	656.57	24.70	546	761
c	100000	772.68	26.61	654	888
D	100000	7426.73	43.76	7223	7622
d_1	100000	0.00	0.00	0	0
d_0	100000	0.00	0.00	0	0
d	100000	0.00	0.00	1E-10	1E-10
Sample size	100000	1545.23	36.19	1394	1708

Test Statistics	Number of Simulations	Mean	Std Dev	Min	Max	Type-I error
Test of difference in SE (McNemar)	100000	0.90	1.33	0	15.68	4.29%
Test of difference in SP (McNemar)	100000	0.95	1.37	0	20.01	4.54%
Test of difference in SE (Wald test)	100000	1.00	1.41	8E-12	16.32	5.00%
Test of difference in SP (Wald test)	100000	1.02	1.44	2E-11	23.33	5.25%

Figure 1. Power vs. Difference in Sensitivity for McNemar's test and Wald Test (bigV) under different sampling schemes (Equally sampling from A, B, C, D cells; Equally sampling from A, B, C cells; Equally sampling from B, C cells) with Prevalence of GDM=0.05.

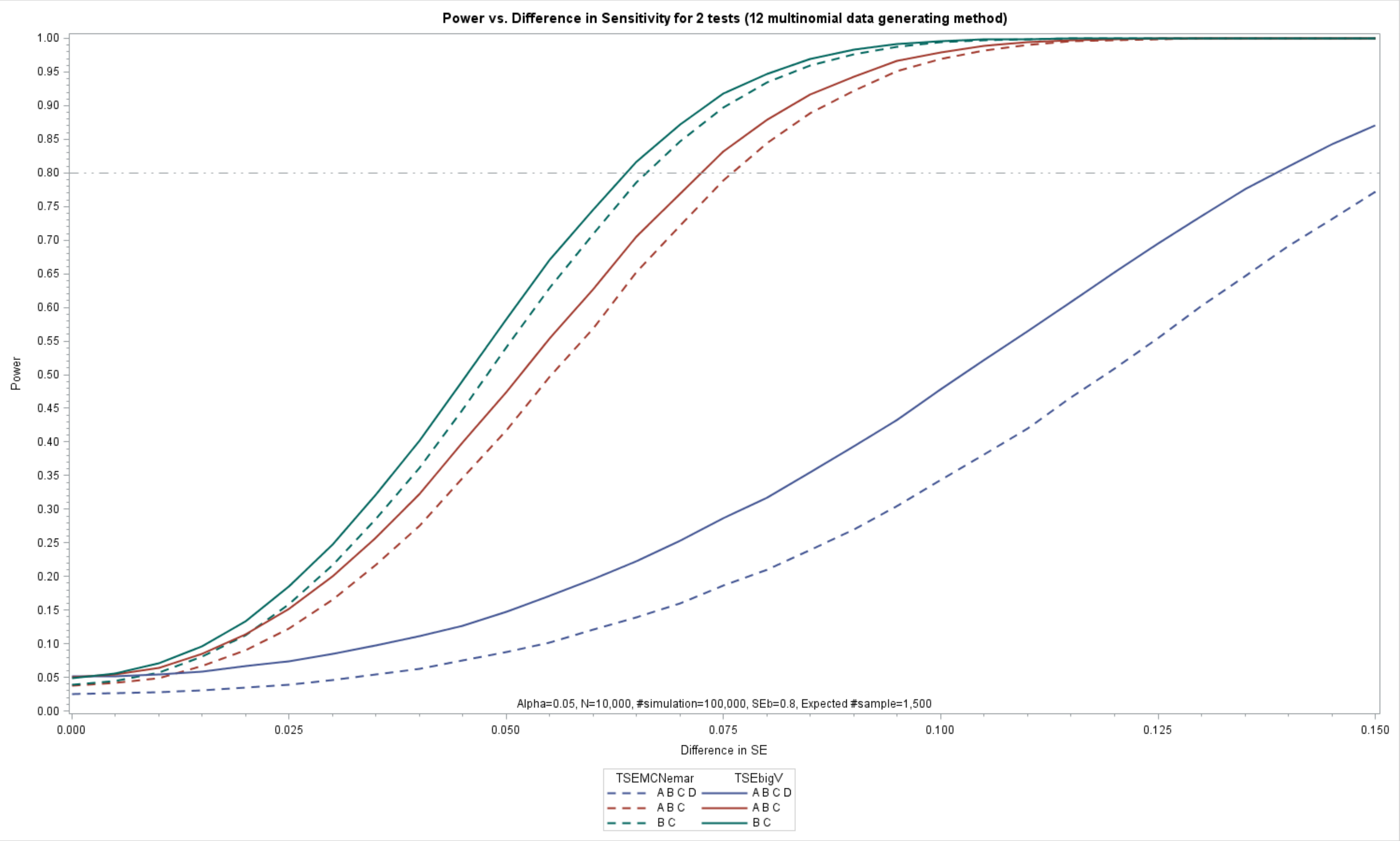


Figure 2. Power vs. Difference in Sensitivity for McNemar’s test and Wald Test (bigV) under different sampling schemes (Equally sampling from A, B, C, D cells; Equally sampling from A, B, C cells; Equally sampling from B, C cells) with Prevalence of GDM=0.10.

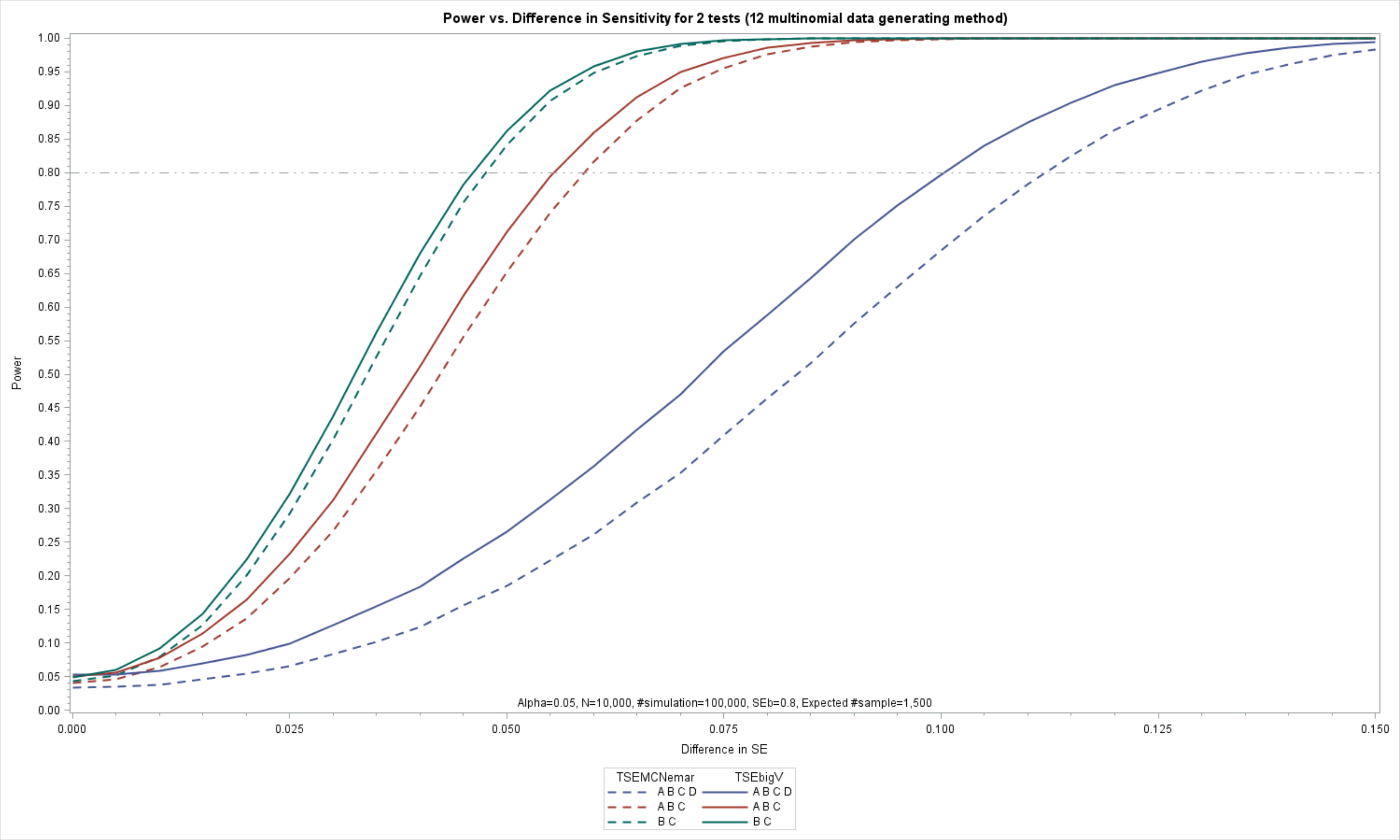


Figure 3. Power vs. Difference in Sensitivity for McNemar's test and Wald Test (bigV) under different sampling schemes (Equally sampling from A, B, C, D cells; Equally sampling from A, B, C cells; Equally sampling from B, C cells) with Prevalence of GDM=0.20.

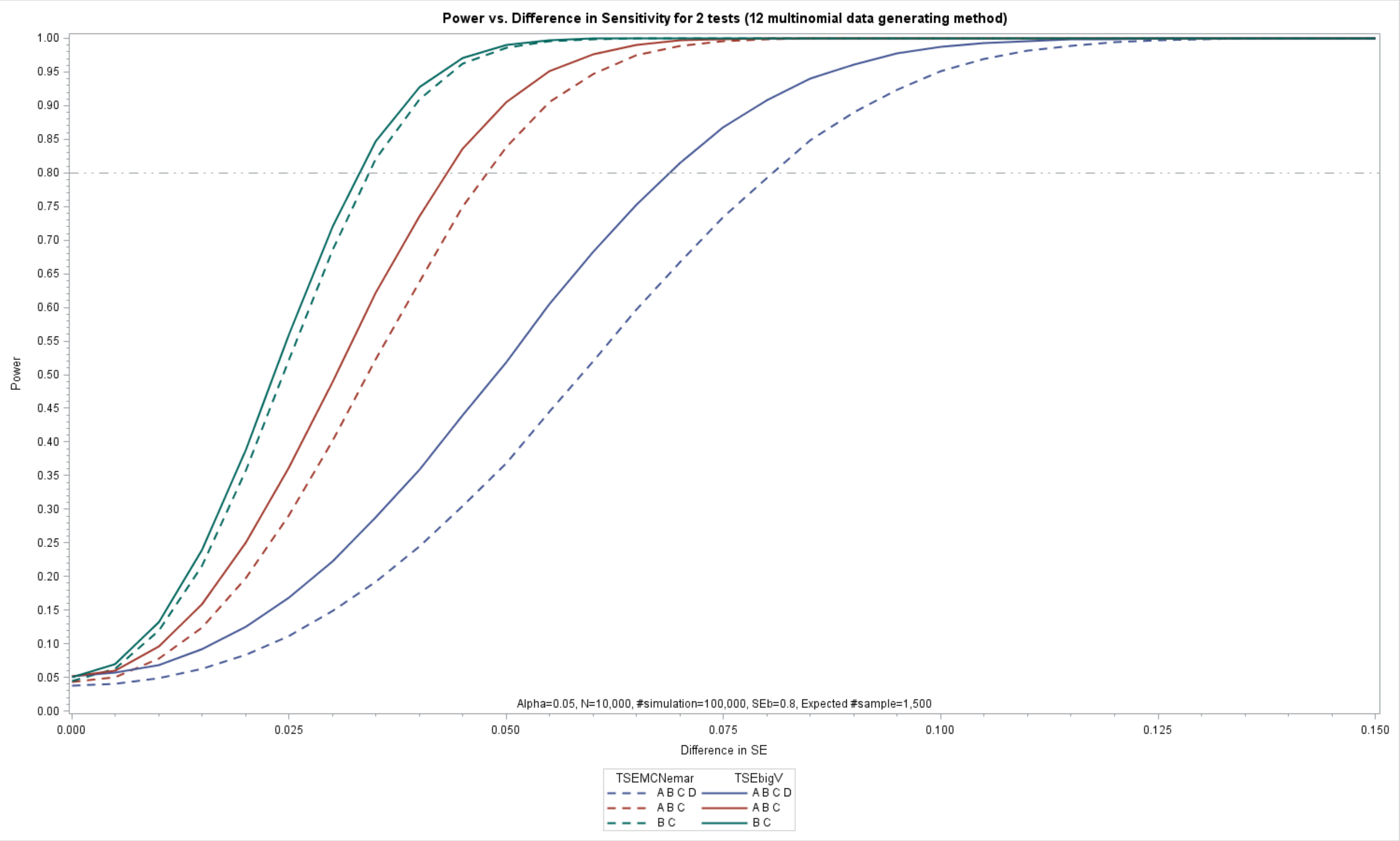


Figure 4. Power vs. Difference in Sensitivity for McNemar's test and Wald Test (bigV) under different sampling schemes (Equally sampling from A, B, C, D cells; Equally sampling from A, B, C cells; Equally sampling from B, C cells) with Prevalence of GDM=0.40.

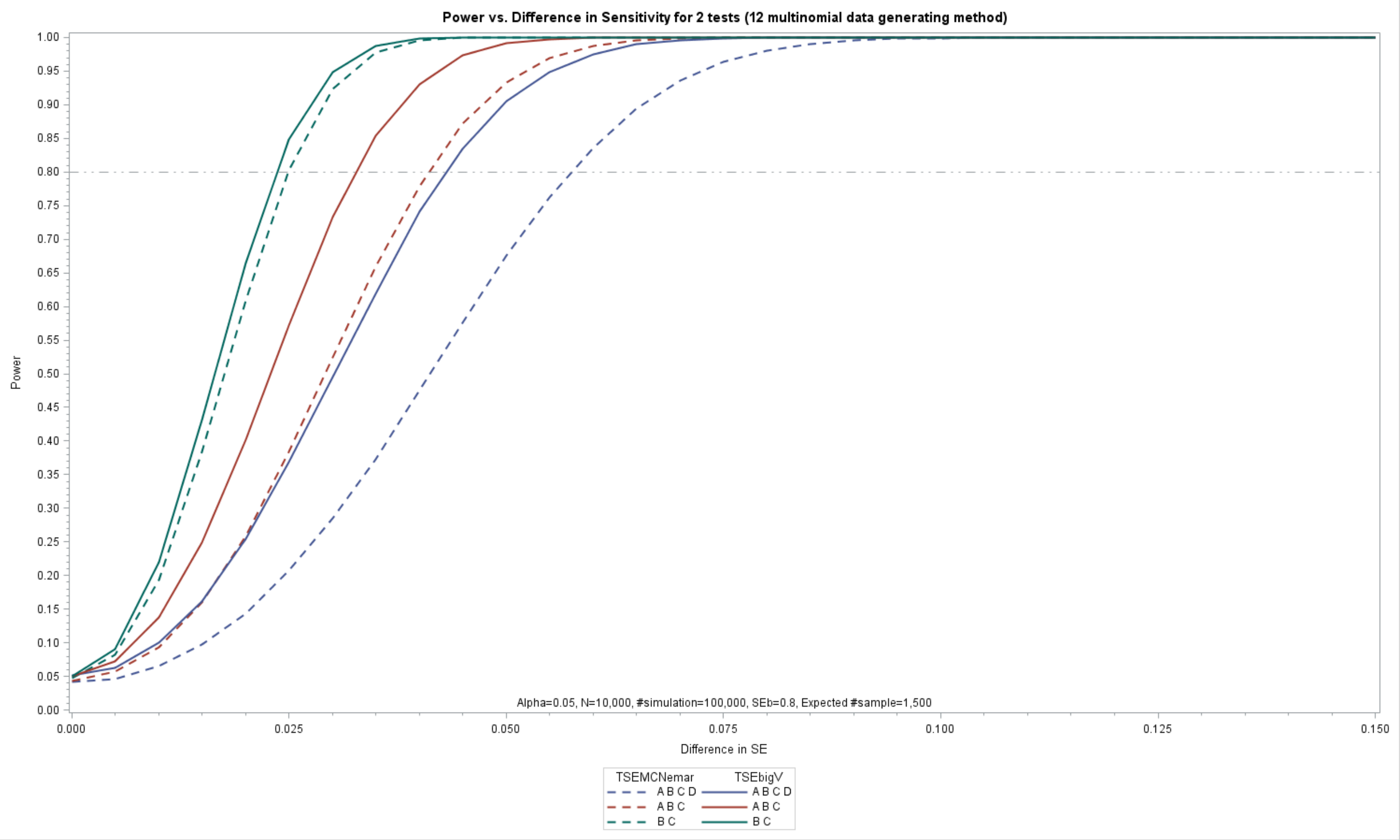


Figure 5. Power vs. Difference in Specificity for McNemar's test and Wald Test (bigV) under different sampling schemes (Equally sampling from A, B, C, D cells; Equally sampling from A, B, C cells; Equally sampling from B, C cells) with Prevalence of GDM=0.05.

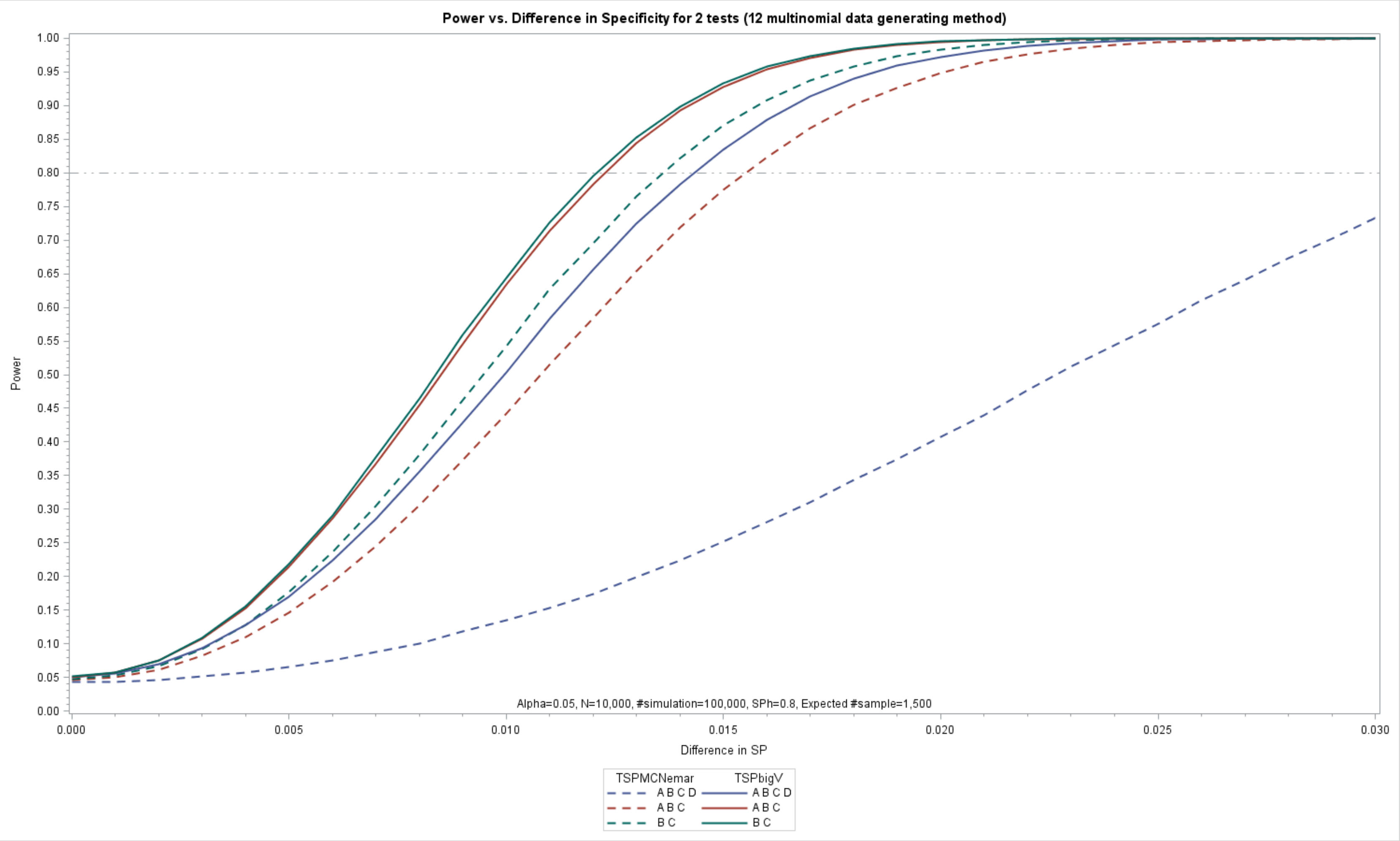


Figure 6. Power vs. Difference in Specificity for McNemar's test and Wald Test (bigV) under different sampling schemes (Equally sampling from A, B, C, D cells; Equally sampling from A, B, C cells; Equally sampling from B, C cells) with Prevalence of GDM=0.10.

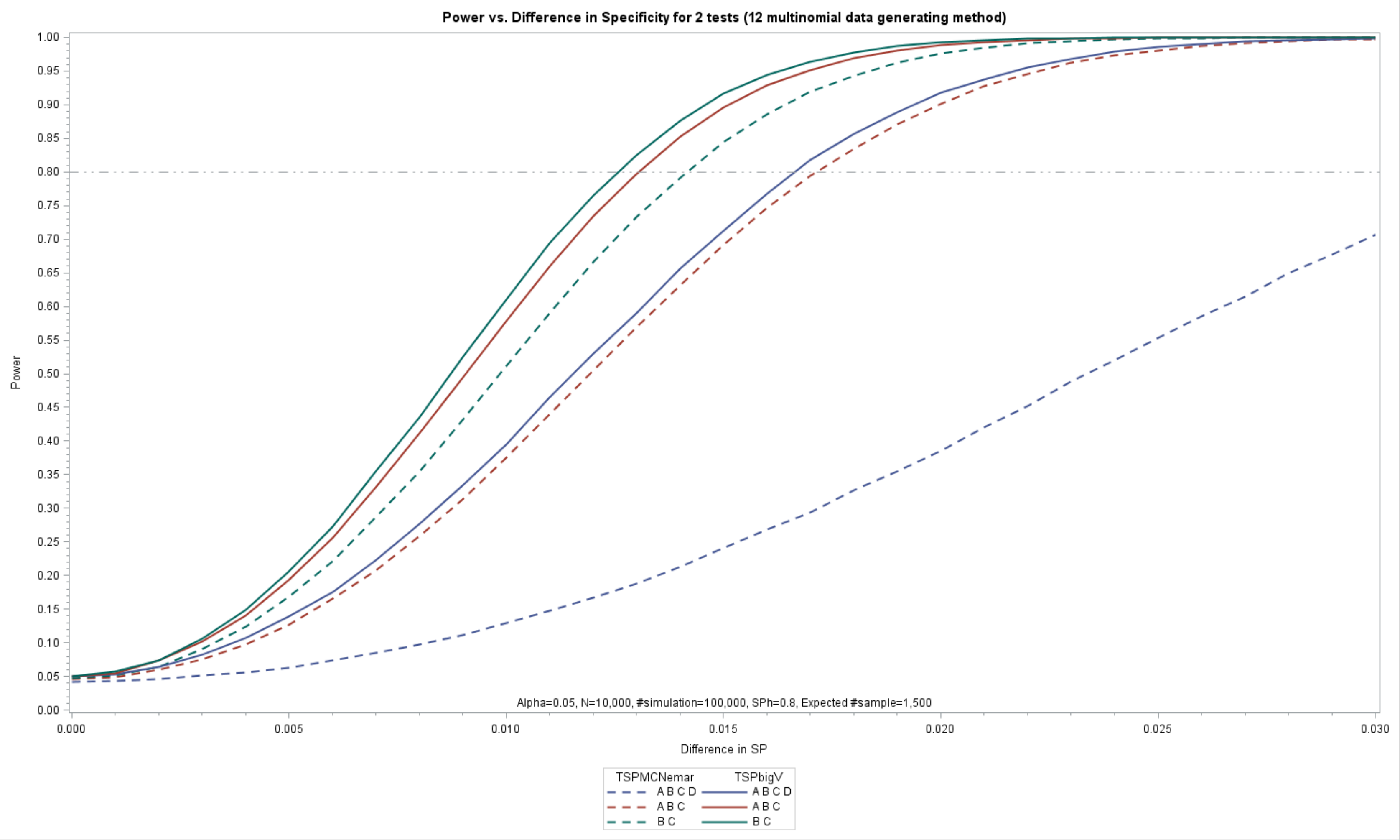


Figure 7. Power vs. Difference in Specificity for McNemar's test and Wald Test (bigV) under different sampling schemes (Equally sampling from A, B, C, D cells; Equally sampling from A, B, C cells; Equally sampling from B, C cells) with Prevalence of GDM=0.20.

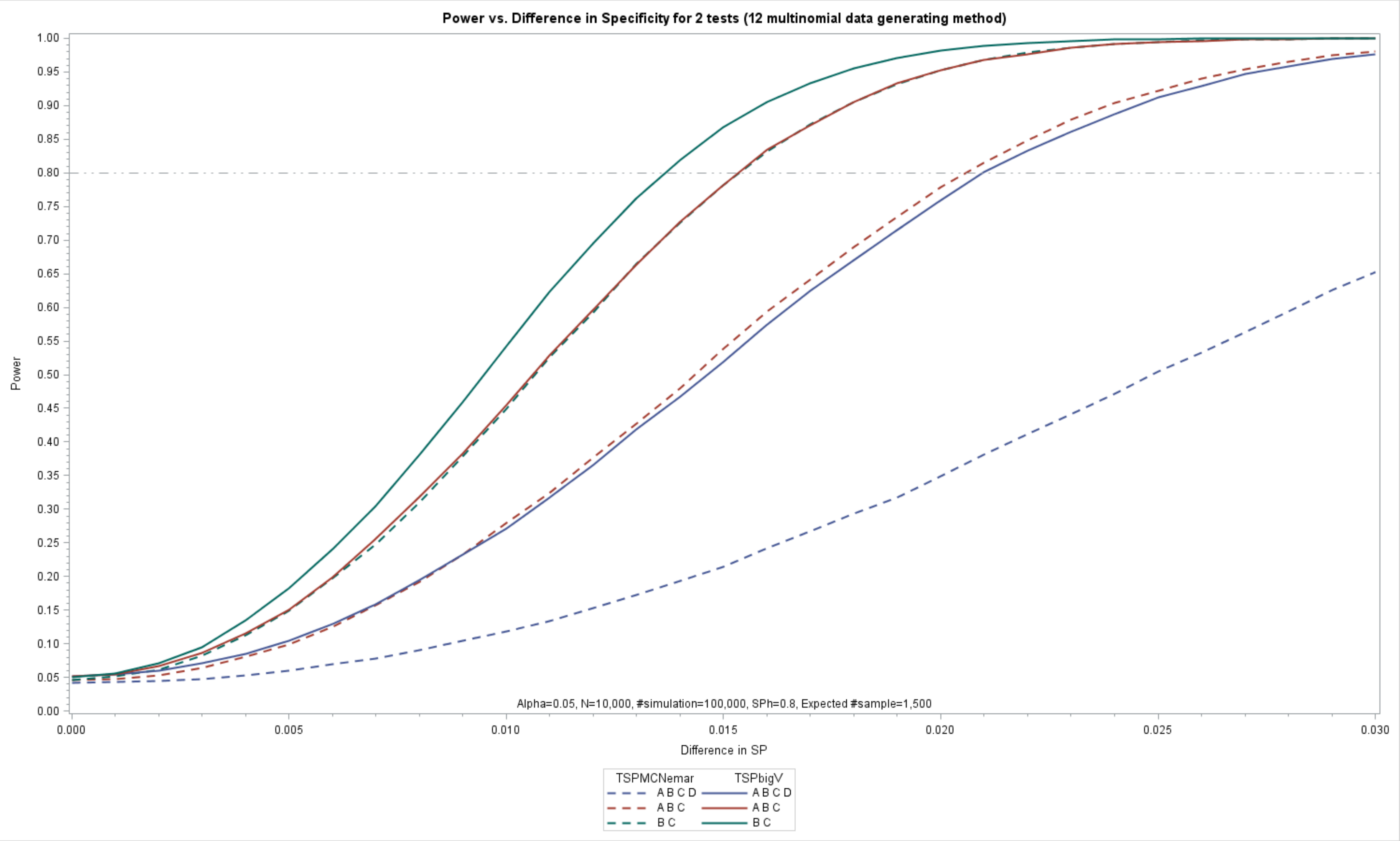
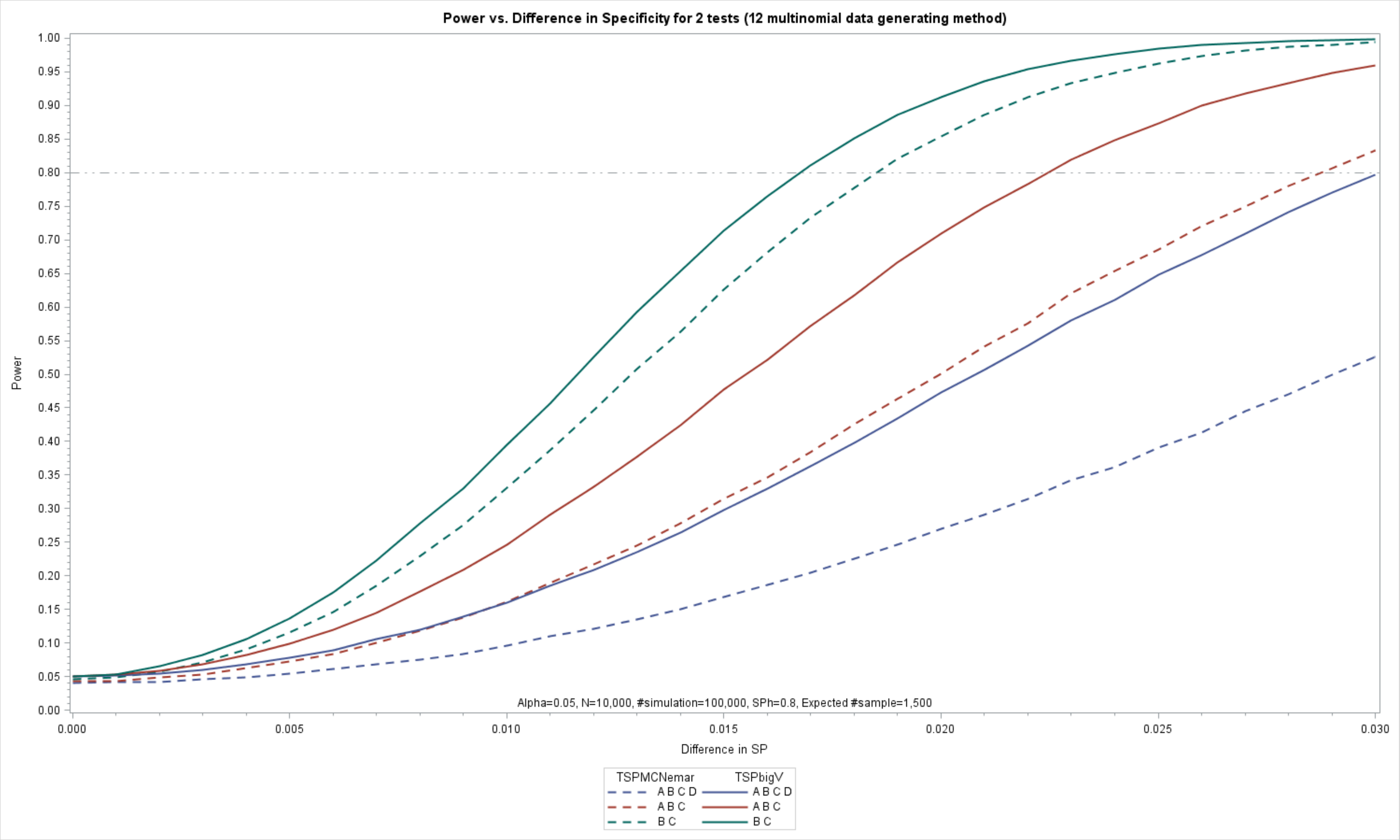


Figure 8. Power vs. Difference in Specificity for McNemar's test and Wald Test (bigV) under different sampling schemes (Equally sampling from A, B, C, D cells; Equally sampling from A, B, C cells; Equally sampling from B, C cells) with Prevalence of GDM=0.40.



Discussion

This study concerns efficient sampling allocations and proposes a statistical method for testing differences in accuracy for two diagnostic methods of a disease. As noted in Results, instead of using random A, B, C, D cells sampling or A, B, C cells sampling along with a conventional McNemar’s test, we recommend to use only B, C cells sampling to get validation sample and to use the Wald test we proposed to test the difference in both sensitivity and specificity of two diagnostic tests (BC and HD). This is because the main difference between a conventional McNemar’s test for paired-match data and our proposed Wald test based on a 12-level multinomial distribution is the assumption behind these two tests. The McNemar’s test requires random sampling among subjects with known true condition status and subsequent determination of the outcomes of the two diagnostic tests. In contrast, our proposed Wald test is based on knowing the outcomes of the two tests before sampling subject (potentially non-randomly, for efficiency reasons) for whom the ‘gold standard’ test will be applied. Also, by using only B, C cells sampling, we are getting more ‘useful’ information to do the hypothesis test, and therefore benefit from higher statistical power, given the same or very similar validation sample size n. Another advantage of using the proposed test is that it does not require the sampling rate of B cell and C cell to be equal (i.e., if we want, we can sample 20% of patients in the B cell and 10% in the C cell, and the test still functions well). However, the McNemar’s test approach requires equal sampling of B cell and C cell.

One limitation of only sampling from the B and C cells is while we can construct valid test statistics to compare them, we cannot directly estimate diagnostic test properties such as sensitivity and specificity for BC and HD. Further, for other diagnostic properties such as negative predictive value (NPV) or positive predictive value (PPV) for BC and HD, a restriction to B and C cell sampling allows neither estimation nor a statistical test for comparison. Referring to Table 2, this is because: $PPV_{HD} = \frac{A_1+C_1}{A+C}$, $PPV_{BC} = \frac{A_1+B_1}{A+B}$, $NPV_{HD} = \frac{D_0+B_0}{D+B}$, $NPV_{BC} = \frac{D_0+C_0}{D+C}$. From these equations, we know we cannot estimate the difference of PPVs without observing patients from the A cell, and likewise, we cannot estimate the difference of NPVs without observing patients from the D cell. Thus, to estimate the differences of NPV and PPV, we would need to do A cell and D cell sampling as well. So one of the 'costs' of this B cell and C cell validation sampling is that we give up doing any test regarding the difference of NPVs and PPVs.

Another aspect of this study to be aware of is that N is not the size of the true population. As noted in Introduction and Methods, we did not consider the finite population setting by which all members of a population have BC and HD data; rather we consider that to be a large overall sample of size N. Because not all members of the population have both BC and HD data, we hereby only focus on a subset of the population whose BC and HD data are both accessible. If the whole population were assessed for BC and HD, finite population corrections may need to be applied.

Besides the Wald test, there exist two classic alternative tests, namely the Score test and the Likelihood-Ratio test. We can also construct a Score test or a Likelihood-Ratio test based on the maximum likelihood estimator $\widehat{D}_{SE}$ and $\widehat{D}_{SP}$ noted in the Methods section and their corresponding likelihood functions. The three tests would be asymptotically equivalent, but there may be advantages to the Wald test in terms of performance and/or ease of use for, when we have a very large overall sample of size N and validation sample of size n (which means we used an underlying assumption: $N \to \infty$ and $n \to \infty$) [**21-22**]. But in moderate or small-sized samples, the Wald test can be extremely conservative when truth is far from the null hypothesis. In this case, an exact test is more desirable (with very small sample size), or the Likelihood-Ratio test (with moderate sample size) may be a better option.

Also, in this study, we set the sampling rates from A, B, C, D cells as $\varphi_A \sim \varphi_D$ so that the expected validation sample size $E(n)$ is very close to 1500, to make different validation sampling allocations comparable. The real-world sampling scheme that applies directly to this 12-level multinomial mechanism would be to apply Bernoulli sampling for each subject in the A-D cells, according to the specified sampling rate. However, another intuitive way to envision the sampling is to specify the actual sampling fractions to be used, so that repeated replications of the experiment would yield the same observed fraction a/A, b/B, etc. Another possible way would be to set the exact numbers of patients to be sampled from A, B, C, D cells upon replications of the experiment, instead of the sampling rates or fractions in each cell. This latter approach would yield a fixed number for the total sample size n, with $Var(n) = 0$. But this method of envisioning the sampling could lead to another issue of 'oversampling', where the generated total number of patients in the A, B, C, D cells (A, B, C, D) are smaller than the number of patients (a, b, c, d) specified to be sampled from each cell. Fortunately, our empirical studies show that the proposed testing approach based on the 12-level multinomial model is valid for use under any of the three ways of envisioning the real-world sampling. This should be a comfort to those seeking to apply these methods in practice for comparing the diagnostic properties of two tests.

References

[1] https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes/gestational

[2] Metzger BE, Coustan DR. (1998) Proceedings of the Fourth International Workshop-Conference on Gestational Diabetes Mellitus. *Diabetes Care*. 21(Suppl. 2): B1–B167

[3] AMERICAN DIABETES ASSOCIATION. (2003) Gestational Diabetes Mellitus. *Diabetes Care*, 26(Suppl. 1): S103-S105.

[4] Carr DB, Gabbe S. (1998). Gestational Diabetes: Detection, Management, and Implications. *Clinical Diabetes*. 16(1): 4.

[5] Anne V, and Isabelle F. (2011). Consequences of gestational and pregestational diabetes on placental function and birth weight. *World J Diabetes*. 2(11): 196–203.

[6] Wendland EM, et al. (2012). Gestational diabetes and pregnancy outcomes--a systematic review of the World Health Organization (WHO) and the International Association of Diabetes in Pregnancy Study Groups (IADPSG) diagnostic criteria. *BMC Pregnancy Childbirth*. 12: 23.

[7] Wong T, Ross GP. (2013). The clinical significance of overt diabetes in pregnancy. *Diabet Med*. 30(4): 468-474.

[8] Chen Y, et al. (2009) Cost of gestational diabetes mellitus in the United States in 2007. *Popul Health Manag*. 12(3): 165–174.

[9] Jacinda MN, et al. (2017) Patterns of gestational diabetes diagnosis inside and outside of clinical guidelines. *BMC Pregnancy and Childbirth.*17: 11

[10] National Diabetes Data Group. (1995) Diabetes in America, 2nd ed. Bethesda, MD: National Institutes of Health.

[11] HAPO Study Cooperative Research Group, Metzger B, Lowe L, et al. (2008) Hyperglycemia and adverse pregnancy outcomes. *N Engl J Med*. 358(19): 1991-2002.

[12] American Diabetes Association. (2012) Position statement: standards of medical care in diabetes. *Diabetes Care*. 35(Suppl. 1): S11-S63.

[13] Carla LD, Shin YK. (2014). Prevalence Estimates of Gestational Diabetes Mellitus in the United States, Pregnancy Risk Assessment Monitoring System (PRAMS) 2007-2010. *Preventing Chronic Disease*. 11: E104.

[14] https://www26.state.nj.us/doh-shad/query/UBQueryTechNotes.html

[15] https://www.cdc.gov/nchs/nvss/vital_certificate_revisions.htm

[16] https://www.cdc.gov/nchs/data/dvs/panelreport_acc.pdf

[17] Dietz PM, Vesco KK, et al. (2008). Postpartum screening for diabetes after a gestational diabetes mellitus-affected pregnancy. *Obstet Gynecol*. 112: 868-874.

[18] Hillier TA, et al. (2013). Markedly different rates of incident insulin treatment based on universal gestational diabetes mellitus screening in a diverse HMO population. *Am J Obstet Gynecol*. 209: 440. E441-449.

[19] Bernard R. (2010). Fundamentals of Biostatistics 7th ed. P375-376.

[20] SAS/STAT 9.3 User's Guide (2011) retrieved on Mar.03 2017 at: https://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#titlepage.htm

[21] Jack J, John D. (2001). Econometric Methods 4th ed. P150.

[22] Thomas P. (2017). Finite Sample Distributions of the Wald, Likelihood Ratio and Lagrange Multiplier Test Statistics in the Classical Linear Model. *Communications in Statistics - Theory and Methods*. 11: p5195-5202