2017

In Search of the Criterion Standard Test in Diagnostic Testing

Thanos D. Kantarelis
Managing Consultant, Navigant

Demetri Kantarelis
Assumption College, dkantar@assumption.edu

Addleton Academic Publishers

American Journal of Medical Research 4(1), 2017
pp. 118–140, ISSN 2334-4814, eISSN 2376-4481 doi:10.22381/AJMR4120179

IN SEARCH OF THE CRITERION STANDARD TEST IN DIAGNOSTIC TESTING

THANOS D. KANTARELIS
Managing Consultant, Navigant,
Chicago, IL
DEMETRI KANTARELIS
dkantarelis@post.harvard.edu
Assumption College, Worcester, MA
(corresponding author)

ABSTRACT. Given a certain technology or procedure for diagnostic testing, different cutoff points produce different sensitivity and specificity rates. The cutoff point that generates highest sensitivity and specificity establishes the Criterion Standard Test (otherwise known as the Gold Standard Test). If, subject to good reason, a new testing technology or procedure emerges, the optimum cutoff point associated with it may generate higher sensitivity and specificity and thus a new improved Criterion Standard Test. Various cutoff selection methodologies have been proposed, all based on Euclidean geometry, involving the so-called Receiver Operating Characteristic (ROC) curve. Our purpose in this paper is to recommend a new selection methodology based on the P-Value associated with the well-known Pearson's chi-squared test (χ^2) – the conventional test utilized when testing for dependence between state of nature (disease present or not present) and evidence (test positive or negative measures). Using a hypothetical numerical example, we demonstrate that the cutoff point associated with the lowest P-Value of the Pearson's chi-squared test is the one that maximizes sensitivity and specificity, or overall accuracy, thus establishing the Criterion Standard Test. Although the best geometric method (sums of squares) and the proposed method are equally effective in selecting the optimum cutoff point, only the proposed new procedure selects based on statistical significance. Additionally, we propose a simple theoretical benefits / costs linear setting to discuss the importance of net benefits associated with testing accuracy and reference harmful as well as beneficial testing cases found in various literature sources.

Keywords: diagnostic testing; criterion standard test; statistics; receiver operating characteristic; FDA; net economic benefits

How to cite: Kantarelis, Thanos D., and Demetri Kantarelis (2017), "In Search of the Criterion Standard Test in Diagnostic Testing," *American Journal of Medical Research* 4(1): 118–140.

Received 23 January 2017 • Received in revised form 3 March 2017
Accepted 4 March 2017 • Available online 20 March 2017

1. Introduction

Epidemiological studies, genetic theory, clinical studies, and testing for efficacy of new medicine and medical devices or procedures, enable researchers and regulatory authorities to estimate probabilities in their efforts to deal with the diagnosis and cure of a disease or, alternatively stated, to minimize false-positives (F+) and false-negatives (F-) that impose costs on society including poor medical outcomes, direct costs associated with less efficient care, inappropriate use of therapies and diagnostic tests, lost patient productivity (e.g., increased absenteeism), and administrative burden.

For example, an epidemiological study establishes state of nature probabilities (or prior probabilities such as the prevalence of a disease in a human population) against which a researcher may test the efficacy of a new medicine or the sensitivity of new diagnostic test or medical device / procedure. Similarly, after genetic theory (e.g., applied to autosomal recessive diseases) establishes prior probabilities, whether an individual carries a disease may be tested subject to optimal cutoff points (cutoff points that maximize test accuracy). Also, clinical studies enable researchers to test their hypotheses (e.g., how likely it is that, given symptoms, a patient carries a disease) based on prior probabilities derived from literature and their clinical experience. Likewise, the U.S. Food and Drug Administration (FDA) requires that the safety of food and cosmetics, and the safety and efficacy of drugs and medical devices are tested and validated or demonstrated.[1]

Recent epidemiological studies have dealt with physicality of older women in Scotland (Yang et al., 2017), children with disorders (Katusic et al., 2017), ageing (Lu et al., 2016), and immunology (Black at al., 2016) as well as mind-body therapy (Bower et al., 2016).

Additionally, clinical research efforts, facilitated by the FDA, have been producing safer, faster and more effective outcomes; most notably, see Zarin et al. (2016) on trial reporting, Schwartz et al. (2016) on new drug approval, Bourgeois et al. (2016) on intervention trials, Russek-Cohen et al. (2011) on diagnostic devices, and Ziegler et al. (2005) on radiology technologies. Undoubtedly, the research effort has been aided by the digital revolution which has greatly contributed to improved diagnostic accuracy and screening; see, among many others, Willis et al. (2011), Albert (2009), Zhou et al. (2011), Zou et al. (2011), and Ballard-Barbash et al. (1997).

Moreover, genomic testing studies have been pushing the evolutionary frontier across the board; specifically, studies by Nair et al. (2016) on endometrial cancer, Stranneheim et al. (2016) on monogenic disorders, Van

Driest et al. (2016) on arrhythmia, Gonzaga-Jauregui et al. (2012) on general lessons associated with human genome sequencing in health and disease, Gepts (2014) on genetic and genomic approaches to plant domestication studies, and Manrai et al. (2016) on genetic misdiagnoses.

Many researches involved in such studies rely on the Bayesian Theorem to derive posterior probabilities based on prior probabilities. According to Copi et al. (2007), Thomas Bayes was "the first to use probability inductively and who established a mathematical basis for probability inference: a means of calculating, from the frequency with which an event has occurred in prior trials, the probability that it will occur in future trials."[2]

Bayesian learning starts with some initial information about an event X which enables the researcher to estimate the probability of event X occurring; in turn, in the next period, if additional or better information becomes available a new probability is estimated (the posterior probability) given the probability estimated in the previous period (the prior probability) and so forth for any n periods. In every new period a new posterior probability is estimated which becomes the prior probability in the next period. Hence, since the posterior probability is based on more and / or better information, it contributes to more and / or better knowledge; it takes us closer to the truth but inductively so: the process generates a probable credible result but not a certain one. Flow Chart 1 sketches this process.[3]

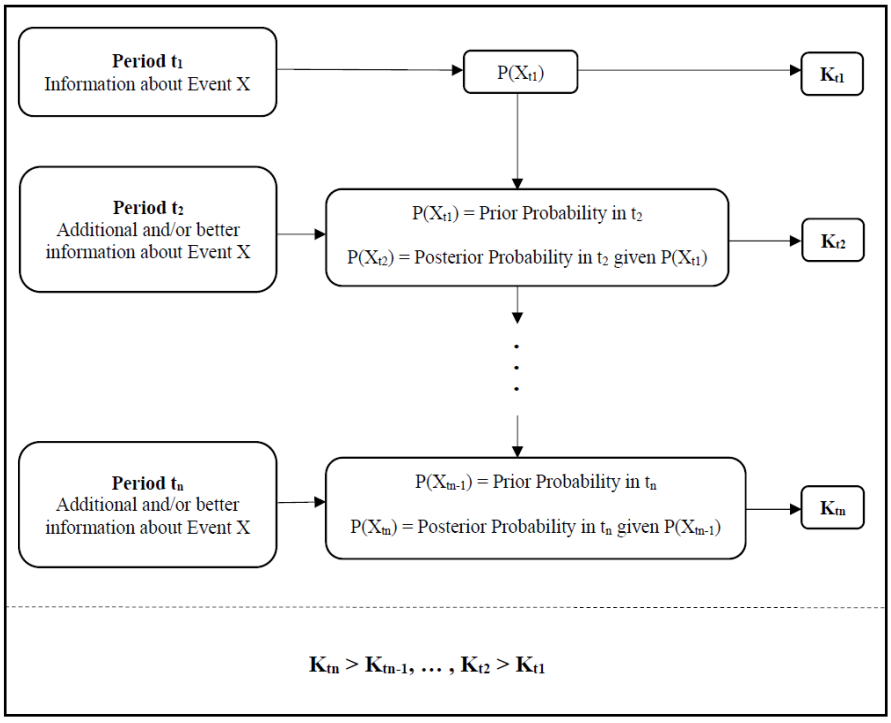
When evaluating a diagnostic test or procedure, different cutoff points produce different sensitivity and specificity rates. The cutoff point that generates the highest sensitivity and specificity establishes the Criterion Standard Test (otherwise known as Gold Standard Test). Of course, if a new diagnostic test or procedure emerges, the optimum cutoff point associated with it may generate higher sensitivity and specificity and thus a new improved Criterion Standard Test. It is also likely that a new testing technology or procedure generates reduced accuracy in which case we revert to the previous Criterion Standard Test.

Various cutoff selection methodologies have been proposed, all based on Euclidean geometry, involving the so-called Receiver Operating Characteristic (ROC) curve. Our purpose in this paper is to recommend a new selection methodology based on the P-Value associated with the well-known Pearson's chi-squared test (χ^2) when testing for dependence between state of nature (disease present or not present) and evidence (test positive or negative measures). Using a numerical example, we shall attempt to demonstrate that the cutoff point associated with the lowest P-Value of the Pearson's chi-squared test is the one that maximizes sensitivity and specificity, or overall accuracy, thus establishing the Criterion Standard Test.

We proceed as follows: in Section 2, we review the existing cutoff methodologies. In Section 3, we offer a hypothetical numerical example that

involves diagnosing cancer with positron emission tomography (PET) and the measure it produces called standardized uptake value (SUV) – an indicator of how likely the part of the body contains cancerous cells. In Section 4, we take the opportunity to discuss some Criterion Standard Test applications found in the literature and we stress the importance of false-positives and false-negatives as costs to society in the discovery process for new diagnostic test / procedure and medicine. Finally, in Section 5 we summarize and conclude. Appendix 1 describes the Bayesian Theorem (statement, proof and examples) which may be skipped by readers familiar with it. All hypothetical data used in Section 3 is in Appendix 2.

Flow Chart 1 Bayesian Learning
(P = Probability, K = Knowledge, t = time ranging from 1 to n)



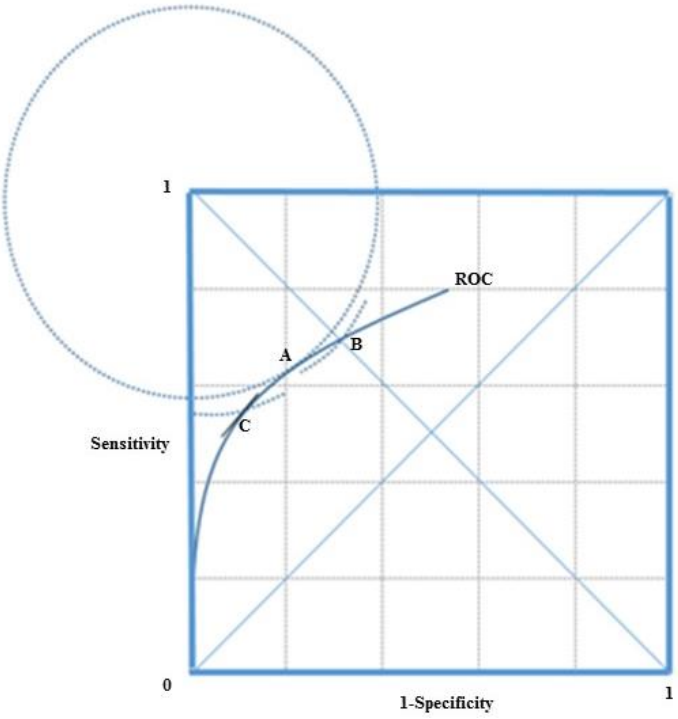
2. Cutoff Methodologies

Existing cutoff methodologies are eloquently described by Froud and Abel (2014) in conjunction with the well-known Receiver Operator Characteristic (ROC) curve which maps "Sensitivity" (vertical axis) vs. "1-Specificity" (horizontal axis).[4] They propose a methodology for the identification of

optimal cutoff points which outperforms the Farrar method and the EMGO method. In their words, "to identify Minimally Important Change (MIC) thresholds on scales that measure a change in health status … we choose the point in ROC space that minimizes [Q_{FA}] … the sums of squares of 1-sens and 1-spec" where Q_{FA} = min $\{(1\text{-Sensitivity})^2 + (1\text{-Specificity})^2\}$.

Assuming that three possible points on the ROC are A, B, and C, Figure 1 describes the Euclidean geometry associated with the Froud-Abel selection result (point A) relative to Farrar (point B) and EMGO (point C). Since the objective is to select the point on the ROC closest to (0,1), or closest to the northwest point, clearly the Froud-Abel method outperfoms the other two. (The circle, or the equidistant frontier $\bar{Q}_{FA}$, is centered around the top-left corner; the equidistant frontier passing through A is closer to the top-left corner relative to the frontiers that pass through B and C). For a different approcah on how to search for optimal cutoff points see Terluin et al. (2015).

Figure 1 The Froud-Abel Method
[Euclidean geometry associated with the Froud-Abel selection result (point A) relative to Farrar (point B) and EMGO (point C)]



Alternatively, as we propose in this paper and show below by way of hypothetical example, the cutoff point that establishes the Criterion Standard

Test is the one that corresponds to the lowest P-Value of the Pearson's chi-squared test (χ^2). More specifically, when testing for the existence of dependence between the state of nature variable (disease present or not present) and the hypothesis or evidence variable (test positive or negative measures) using the χ^2 test, in most applications, we end up with many cutoff points that enable us to reject the zero hypothesis H_0 (the state of nature variable and the hypothesis variable are independent) in favor of the alternative hypothesis H_1 (the state of nature variable and the hypothesis variable are not independent); hence, which one of the many cutoff points that generate statistically significant results should be selected? We propose that the cutoff point that generates the lowest P-Value ought to be considered as the Criterion Standard Test. In conventional research, the gold standard is reported void of statistical significance; as such, it is less useful in decision making involving screening and diagnostic testing or in the process of medicine discovery. The proposed methodology adds statistical significance to the process of establishing maximum accuracy (or a gold standard) thus making decisions more credible. Using a fictitious numerical example, we show below that the lowest P-Value of the χ^2 test corresponds to the cutoff point identified by the Froud-Abel Method as well.

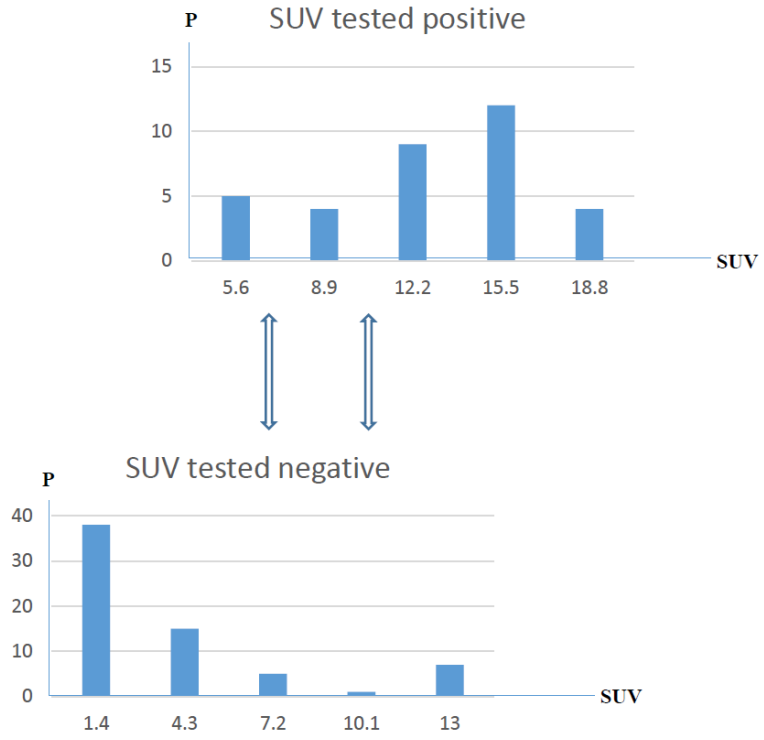
3. Hypothetical Numerical Example

Positron emission tomography (PET), among other applications, may be used to diagnose cancer. PET generates a *standardized uptake value* (SUV) which serves as an indicator of the likelihood of cancer. SUV is a positive number ranging from 0 upwards; the higher the SUV value the more likely it is that cancer is present. A value greater than 10 implies a high likelihood for aggressive disease. After SUV is measured the patients undergo a biopsy wherein a small piece of tissue from the suspected area is removed and examined histologically and/or genetically sequenced to inform a cancer diagnosis. Pathological verification along these lines gives rise to the so-called gold standard.

In the table that appears in Appendix 2 we report hypothetical data for 100 individuals: first column – identification of subjects (ID), second column – SUV scores, and third column – biopsy results for Cancer where 1 = present and 2 = not present. Figure 2 reports the sample probability distributions of SUV tested positive (top) and tested negative (bottom); it shows that higher SUV scores are more likely to be associated with cancer than otherwise and it clearly demonstrates the impact of cutoff point regarding false-positives and false-negatives: when the cutoff point is increased from the left double-headed arrow to the right double-headed arrow, false-positives decrease and false-negatives increase.[5] For "treatable" cancer, a test that generates a high

number of false-negatives is not a good test and undoubtedly more problematic than a test that generates a high number of false-positives. False-positives would cause psychological discomfort and unnecessary treatment, sometimes even surgery or chemotherapy, but, false-negatives, may delay treatment and could lead to loss of life. On the other hand, if the disease is incurable a false-negative diagnosis may not be that bad. Hence, at least for treatable diseases, quickly identifying optimum cutoff points is of paramount importance.

Figure 2 The need for optimum cutoff point
(P = Probability, SUV = Standardized Uptake Value)



To discover the optimum cutoff point (or, the cutoff point that would give rise to the Criterion Standard Test), we proceeded as follows: based on the distribution of SUV scores, we constructed all possible 2-variable contingency tables per SUV value – as the one below for SUV = 10 – and, using the χ^2 test, we tested whether or not the state of nature variable and the test results variable are independent. (Details regarding such contingency tables may be found in Appendix 1.)

State of Nature Variable
(D = Disease present, ND = Disease not present)

Test Variable		D	ND
	Test Positive	TP	FP
	Test Negative	FN	TN

TP = True-Positive, TN = True-Negative,
FP = False-Positive, FN = False-Negative

State of Nature Variable
(C = Cancer present, ND = Cancer not present)

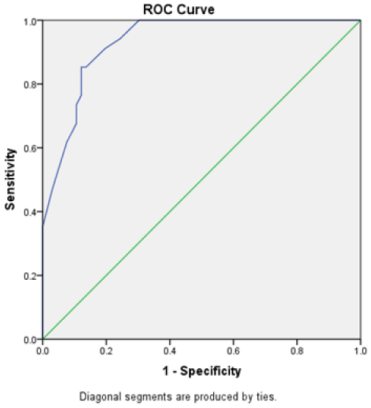
Test Variable		C	NC
	Test Positive: SUV ≥ 10	26	8
	Test Negative: SUV < 10	8	58

For 20 SUV discrete scores, Table 1 reports the following: SUV score and corresponding Sensitivity, Specificity, the Froud-Abel sums of squares (Q_{FA}) result, and the P-Value of each test based on the χ^2. Figure 3 reports the corresponding ROC line. The results show that the optimum cutoff point is SUV = 8. This point is picked by the Froud-Abel method (lowest Q_{FA}) as well as by our proposed new method which relies on the χ^2 test and the lowest P-Value associated with it. Although the two methodologies are equally effective in selecting the optimum cutoff point, only the proposed new procedure selects based on statistical significance: it ranks cutoff points according to the P-Value of the χ^2 test and selects the one that corresponds to the lowest P-Value or highest possible level of statistical significance.

Table 1 Criterion Standard Test: Similarity between Froud-Abel and new method

SUV	Sensitivity	Specificity	Q_{FA}	P-Value (χ^2 test)
1	1	0.13636	0.74587405	0.0225905662698
2	1	0.25760	0.55115776	0.0011245653222
3	1	0.57580	0.17994564	0.0000000200954
4	1	0.69700	0.09180900	0.0000000000389
5	0.9411	0.75760	0.06222697	0.0000000000346
6	0.9118	0.80300	0.04658824	0.0000000000090
7	0.8529	0.86360	0.04024337	0.0000000000027
8	**0.8529**	**0.87880**	**0.03632785**	**0.0000000000007**
9	0.7941	0.87880	0.05708425	0.0000000000234
10	0.7647	0.87880	0.07005553	0.0000000001235
11	0.7353	0.89390	0.08132330	0.0000000001778
12	0.6765	0.89390	0.11590946	0.0000000037146
13	0.6176	0.96970	0.14714785	0.0000000042114
14	0.4706	0.96970	0.28118245	0.0000000523462
15	0.3529	1.00000	0.41873841	0.0000002948080
16	0.2353	1.00000	0.58476609	0.0000361425549
17	0.1765	1.00000	0.67815225	0.0003544775043
18	0.1176	1.00000	0.77862976	0.0053934345806
19	0.0588	1.00000	0.88585744	0.0515240209908
20	0.0294	1.00000	0.94206436	0.1248516808144

Figure 3 ROC



Case Processing Summary

OT	Valid N (listwise)
Positive[a]	34
Negative	66

Larger values of the test result variable(s) indicate stronger evidence for a positive actual state.

a. The positive actual state is 1.00.

Area Under the Curve

Test Result Variable(s): SUV

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.932	.023	.000	.887	.978

The test result variable(s): SUV has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

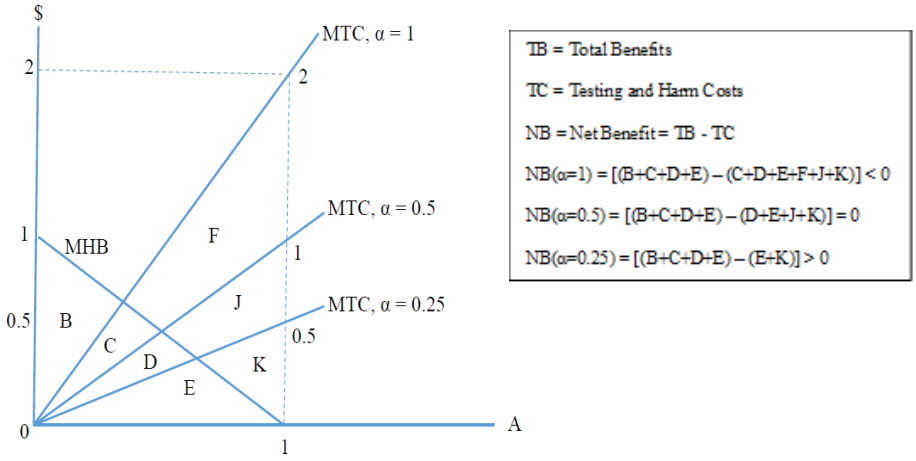
4. Net Benefits of "Accuracy"

The pursuit of accuracy (A) from successive diagnostic tests, where A = [(TP + TN) | (TP + TN + FP + FN)], $0 \leq A \leq 1$, offers improved outcomes, in general, health benefits (HB) and, simultaneously, imposes costs to all parties involved such as testing costs and harms (TC). HB and TC depend on many variables: testing procedures (with direct risks – for example, anesthesia; and indirect risks – for example, stress of testing), laboratory procedures regarding handling of samples and quality of test administration, availability of follow-up tests, potential interference with subsequent tests, and other such variables.

Therefore, we believe, it would be logical to express, rather simplistically but without loss of generality, HB and TC as functions of just A such as $HB = A - 1/2A^2$ and $TC = \alpha A^2$, where α = constant. The derivatives of these functions with respect to A are MHB = 1 – A and MTC = 2αA, where MHB stands for Marginal Health Benefits and MTC for Marginal Testing and Harm Costs.

Figure 4 displays, with $ on the vertical axis and A on the horizontal, the MHB curve together with three MTC curves at various values of α where α may be viewed as an indicator of cost sensitivity. Benefits are maximized at A = 1, the point of maximum accuracy.

Figure 4 Marginal Health Benefits and Marginal Testing and Harm Costs



At the point of maximum A (A=1), the entire area under the MHB curve corresponds to total benefits (TB) and the entire area under the MTC curve, given α, corresponds to total testing and harm costs (TC). The net benefit (NB) is greater than zero when $\alpha < 0.5$.

In the pursuit of accuracy improvement, undoubtedly, innovation in testing, new medical devices and new medicine are difficult, complex and delicate processes that require proper incentives, precautions, and effective regulatory directives with the objective to optimize net health benefits. Although estimation of net benefits ought to be undertaken for each individual test as well as for the whole healthcare industry, it is not our objective to do so in this paper; for useful papers along these lines, especially on cost-utility methodology, see Vanhook (2007) and Wright et al. (2012). However, we believe it would be worthwhile to reference some published reports regarding diagnostic testing accuracy and the net benefit implications associated with it.

In a recent report the FDA [Food and Drug Administration (2015)] informs the public about harmful costs due to false-positives and false-negatives as well as due to treatments based on refuted concepts and inaccurate or untrustworthy tests; as reported in the executive summary (p. 2),

> [l]aboratory developed tests (LDTs) serve an increasingly important role in health care today. They also have become significantly more complex and higher risk, with several notable examples of inaccurate tests placing patients at otherwise avoidable risk … [despite the fact that the examined laboratories] follow the minimum requirements of Clinical Laboratory Improvement Amendments (CLIA) …
>
> We examined events involving 20 LDTs that illustrate, in the absence of compliance with FDA requirements, that these products may have caused or have caused actual harm to patients. In some cases, due to false-positive tests, patients were told they have conditions they do not really have, causing unnecessary distress and resulting in unneeded treatment. In other cases, the LDTs were prone to false-negative results, in which patients' life-threatening diseases went undetected. As a result, patients failed to receive effective treatments.
>
> Other LDTs provided information with no proven relevance to the disease or condition for which they are intended for use, while still others are linked to treatments based on disproven scientific concepts. In addition to patient harm, inaccurate or unreliable tests can be costly to society. We estimated these costs, if sufficient data were available.

Summaries of some FDA reports on LDTs are displayed in Table 2 (clinical consequence highlighted). The summaries correspond to a test that produced many false-positives (ovarian cancer – top), a test that produced many false-negatives (breast cancer – bottom), and a test that produced both many false-positives and many false-negatives (prenatal testing – bottom).

Table 2 Three inaccurate tests as reported by the FDA[a]

OvaSure™ Ovarian Cancer Screening Test

Category	LDT Characteristics
LDT Name	OvaSure Screening Test
Description	Blood test on four biomarkers based on initial research in the published literature reporting an association with ovarian cancer
Purpose	Screen for and detect ovarian cancer
Target Population	Women at risk for ovarian cancer
Alternatives	Other biomarkers or physical symptoms
LDT Problem 1	No validation that test predicts or detects ovarian cancer
LDT Problem 2	Inflated PPV claims by the manufacturer, so many patients with a positive test won't have the disease
Clinical Consequence	**Women with false-positive tests may undergo unnecessary surgery to remove healthy ovaries**
Potential Impact of FDA	Assurance the test meets minimum performance standards
Oversight	Evaluation of manufacturer claims
Cost Impact of Inaccuracy	$12,578 per ovary removal after false-positive

[a]Source: Food and Drug Administration (2016) – Office of Public Health Strategy and Analysis Office of the Commissioner, "The Public Health Evidence for FDA Oversight of Laboratory Developed Tests: 20 Case Studies," November 16, 2015.

Oncotype DX HER2 Breast Cancer RT-PCR Test

Category	LDT Characteristics
LDT Name	Oncotype DX HER2 RT-PCR
Description	Rapid PCR test for tumor HER2 receptors
Purpose	Use HER2 receptor level to guide treatment
Target Population	Newly diagnosed Stage I and II breast cancer patients
Alternatives	FDA-approved HER2 receptor tests
LDT Problem 1	Test has poor sensitivity – many tests reported as normal HER2 levels will actually have high HER2 levels
Clinical Consequence	**Patients with false-negative tests won't receive appropriate treatment, and cancer may progress**
Potential Impact of FDA	Assurance the test meets minimum performance standards
Cost Impact of Inaccuracy	$775,278 estimated cost per false-negative case

[a]Source: Food and Drug Administration (2016) – Office of Public Health Strategy and Analysis Office of the Commissioner, "The Public Health Evidence for FDA Oversight of Laboratory Developed Tests: 20 Case Studies," November 16, 2015.

Noninvasive Prenatal Testing (A.K.A. cell-free DNA testing)

Category	LDT Characteristics
LDT Name	Noninvasive prenatal cell-free DNA testing (NIPT, or cfDNA)
Description	Blood test to identify traces of fetal chromosomes in maternal blood
Purpose	To detect a range of fetal chromosomal abnormalities
Target Population	Pregnant women concerned about a fetal chromosomal abnormality
Alternatives	Invasive testing, including amniocentesis and chorionic villi sampling; "quad testing" of multiple substances combined with ultrasound imaging
LDT Problem 1	Lack of clinical validation that tests detect and predict fetal abnormalities at an appropriate rate
LDT Problem 2	Many false-positive results when used in the general population
Clinical Consequence	**Women with false-positive results may abort a normal pregnancy; women with false-negative results may deliver a child with an unanticipated genetic syndrome**
Potential Impact of FDA Oversight	Assurance the test meets minimum performance standards; evaluation of manufacturer claims
Cost Impact of Inaccuracy	Not estimated

[a]Source: Food and Drug Administration (2016) – Office of Public Health Strategy and Analysis Office of the Commissioner, "The Public Health Evidence for FDA Oversight of Laboratory Developed Tests: 20 Case Studies," November 16, 2015.

But, tests are not all harmful. Lewis (2016) describes the immense benefit and the very low cost of an ingenious new diagnostic test for the Zika virus based on CRISPR/Cas9 gene-editing (a new approach in testing which relies on genome tinkering with demonstrated potential to edit DNA in cell lines and embryos, a methodology that has spurred international discussion about ethical, legal and social issues). In Lewis' words,

> [s]cientists have developed a cheap, rapid, paper-based diagnostic test for Zika virus. … [which] takes only two to three hours … Using CRISPR/Cas9 gene-editing, the test is capable of distinguishing between different strains of Zika .... The Cas9 enzyme selectively targets and cleaves DNA synthesized from viral RNA only if a specific sequence is present, rendering it undetectable by the RNA sensor. If the sequence is not present, the DNA is not cleaved and the virus will be detected. Each test costs less than $1, and can be stored at room temperature for up to a year.

The above examples of unsuccessful and successful diagnostic tests imply (a) that the value of a diagnostic test ultimately lies in its effect on patient

outcomes, (b) that a new test should only be introduced into clinical practice if it is more likely that it would contribute to improving health outcomes, and (c) that decision-making regarding a new test ought to involve selecting, from among many competing tests, the one that generates the highest level of accuracy.

5. Summary and Conclusion

In the sections above, we reexamined how the Criterion Standard Test (otherwise known as the Gold Standard Test) is determined in diagnostic testing, and we proposed a new selection methodology based on the P-Value associated with the well-known Pearson's chi-squared test (χ^2), the test used when testing for dependence between state of nature (disease present or not present) and evidence (test positive or negative measures). With the assistance of a hypothetical numerical example, we demonstrated that the cutoff point associated with the lowest P-Value of the Pearson's chi-squared test is the one that maximizes overall accuracy, thus establishing the Criterion Standard Test. Although our methodology and the sums of squares approach are equally effective in selecting the optimum cutoff point, only the proposed new procedure selects based on statistical significance. Additionally, using a simple benefits / costs theoretical linear setting, we discussed the importance of net benefits of testing for accuracy and referenced harmful as well as beneficial diagnostic tests found in various literature sources.

In general, the proposed statistician test may be readily employed in any biomedical testing procedure described by the National Center for Biotechnology Information (2017) and, more specifically, in conjunction with research involving biomarkers, such as estrogen (ER) and progesterone (PR) receptors, in breast cancer (see Varga et al., 2013). Furthermore, it can be added to the arsenal of tools utilized by the FDA as it endeavors to carry its mission, that is, to inform the public about harmful costs due to false-positives and false-negatives as well as due to treatments based on refuted concepts and inaccurate or untrustworthy tests and products.

Concluding, we would like to remark on the gene-editing revolution that our society currently experiences. Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR), as stressed by The Scientist (Custom publishing, October 2016), "is becoming the main procedure to knock-in or knock-out genes or alter genetic sequences. Due to its simplicity, multiplexing capability and reagent availability, researchers are exploring the limits of its capabilities in model systems and for clinical applications. Efficient screening and detection of gene editing events is critical to successfully generating edited cell lines or organisms."

The potential applications for genetic testing are enormous, given that almost every known disease having some aspect that is influenced by, if not directly caused by, metamorphoses in the genome of the organism. Genome tinkering with demonstrated potential to edit DNA in cell lines and embryos is a revolution to reckon with especially because it triggers debates that relate to, as stressed by Niemiec et al. (2016), ethical, legal and social issues.

Acknowledgments
We are grateful to the three anonymous reviewers for critical and constructive comments on how to improve the manuscript; we remain solely responsible for all remaining omissions and errors.

NOTES

1. As reported by UK's organization Health Knowledge (2016), "diagnostic tests are different than screening tests. The primary purpose of screening tests is to detect early disease or risk factors for disease in large numbers of apparently healthy individuals. The purpose of a diagnostic test is to establish the presence (or absence) of disease as a basis for treatment decisions in symptomatic or screen positive individuals (confirmatory test)." Key differences are reported in the following table:

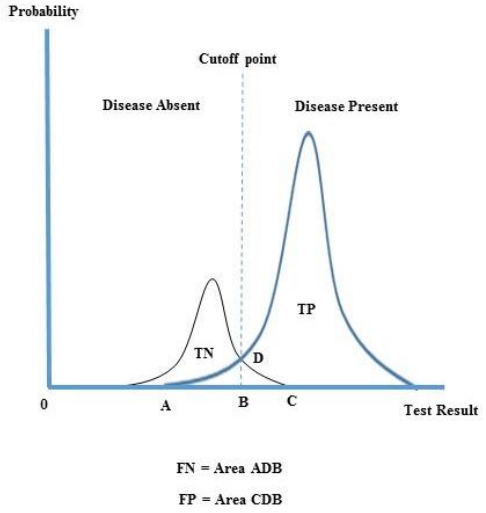
	Screening tests	**Diagnostic tests**
Purpose	To detect potential disease indicators	To establish presence/absence of disease
Target population	Large numbers of asymptomatic, but potentially at risk individuals	Symptomatic individuals to establish diagnosis, or asymptomatic individuals with a positive screening test
Test method	Simple, acceptable to patients and staff	Maybe invasive, expensive but justifiable as necessary to establish diagnosis
Positive result threshold	Generally chosen towards high sensitivity not to miss potential disease	Chosen towards high specificity (true negatives). More weight given to accuracy and precision than to patient acceptability
Positive result	Essentially indicates suspicion of disease (often used in combination with other risk factors) that warrants confirmation	Result provides a definite diagnosis
Cost	Cheap, benefits should justify the costs since large numbers of people will need to be screened to identify a small number of potential cases	Higher costs associated with diagnostic test maybe justified to establish diagnosis.

2. According to Copi et al. (2007), "Inductive reasoning (as opposed to deductive reasoning or abductive reasoning) is reasoning in which the premises are viewed as supplying strong evidence for the truth of the conclusion. While the conclusion of a deductive argument is certain, the truth of the conclusion of an inductive argument is probable, based upon the evidence given." Conventionally, *induction* is reasoning from specific to general and *deduction* is reasoning from general to specific.

3. See Appendix 1 for more details on the Bayesian Theorem.

4. The ROC determines optimal sensitivity and specificity which establishes the highest possible degree of a test's accuracy; and to the extent that diagnostic tests generate different ROCs, the ROC closest to the top left enables us to select the most useful test, the one with even higher "Sensitivity" and lower "1-Specificity."

5. A theoretical depiction of Graph 2, with continuous data and a certain prevalence (e.g., 80%), would look as follows (where TN = Test Negative, TP = Test Positive, FN = False-Negative, FP = False-Positive):



As the graph clearly shows, when the cutoff point moves to the right false-negatives rise and false-positives fall (vice versa when it moves to the left). Thus, balancing this tradeoff between false-negatives and false-positives should be an important goal of diagnostic testing.

REFERENCES

Albert, P. S. (2009). "Estimating Diagnostic Accuracy of Multiple Binary Tests with an Imperfect Reference Standard," *Statistics in Medicine* 28(5): 780–797.

Ballard-Barbash, R., Taplin, S. H., Yankaskas, B. C., Ernster, V. L., Rosenberg, R. D., Carney, P. A., Barlow, W. E., Geller, B. M., Kerlikowske, K., Edwards, B. K., Lynch, C. F., Urban, N., Chrvala, C. A., Key, C. R., Poplack, S. P., Worden, J. K., and Kessler, L. G. (1997). "Breast Cancer Surveillance Consortium: A National Mammography Screening and Outcomes Database," *American Journal of Roentgenology* 169(4): 1001–1008.

Black, D. S., and Slavich, G. M. (2016). "Mindfulness Meditation and the Immune System: A Systematic Review of Randomized Controlled Trials," *Annals of the New York Academy of Sciences* 1373(1): 13–24.

Bourgeois, F. T., Olson, K. L., Tse, T., Ioannidis, J. P., and Mandl, K. D. (2016). "Prevalence and Characteristics of Interventional Trials Conducted Exclusively in Elderly Persons: A Cross-sectional Analysis of Registered Clinical Trials," *PLoS One* 11(5): e0155948.

Bower, J. E., and Irwin, M. R. (2016). "Mind-body Therapies and Control of Inflammatory Biology: A Descriptive Review," *Brain, Behavior, and Immunity* 51: 1–11.

Copi, I. M., Cohen, C., and Flage, D. E. (2007). *Essentials of Logic*. 2nd edn. Upper Saddle River, NJ: Pearson Education.

Encyclopedia Britannica. https://www.britannica.com/biography/Thomas-Bayes

Food and Drug Administration (2016). Office of Public Health Strategy and Analysis Office of the Commissioner, "The Public Health Evidence for FDA Oversight of Laboratory Developed Tests: 20 Case Studies," November 16, 2015.

Froud, R., and Abel, G. (2014). "Using ROC Curves to Choose Minimally Important Change Thresholds when Sensitivity and Specificity Are Valued Equally: The Forgotten Lesson of Pythagoras. Theoretical Considerations and an Example Application of Change in Health Status," *PLoS One* 9(12): e114468.

Gepts, P. (2014). "The Contribution of Genetic and Genomic Approaches to Plant Domestication Studies," *Current Opinion in Plant Biology* 18: 51–59, http://dx.doi.org/10.1016/j.pbi.2014.02.001.

Gonzaga-Jauregui, C., Lupski, J. R., and Gibbs, R. A. (2012). "Human Genome Sequencing in Health and Disease," *Annual Review of Medicine* 63: 35–61.

Health Knowledge (2016). http://www.healthknowledge.org.uk/

Katusic, S. K., Colligan, R. C., Myers, S. M., Voigt R. G., Yoshimasu, K., Stoeckel, R. E., and Weaver, A. L. (2017). "What Can Large Population-based Birth Cohort Study Ask about Past, Present and Future of Children with Disorders of Development, Learning and Behaviour?" *Journal of Epidemiology & Community Health*, DOI: 10.1136/jech-2016-208482

Lewis, T. (2016). "New Zika Diagnostic," *The Scientist*, May 9, http://www.the-scientist.com/?articles.view/articleNo/46052/title/New-Zika-Diagnostic/

Lu, W., Benson, R., Glaser, K., Platts, L. G., Corna, L. M., Worts, D., McDonough, P., Di Gessa, G., Price, D., and Sacker, A. (2016). "Relationship between Employment Histories and Frailty Trajectories in Later Life: Evidence from the English Longitudinal Study of Ageing," *Journal of Epidemiology & Community Health*, DOI: 10.1136/jech-2016-207887.

Manrai, A. K., Funke, B. H., Rehm, H. L., Olesen, M. S., Maron, B. A., Szolovits, P., Margulies, D. M., Loscalzo, J., and Kohane, I. S. (2016). "Genetic Misdiagnoses and the Potential for Health Disparities," *New England Journal of Medicine* 375: 655–665.

Nair, N., Camacho-Vanegas, O., Rykunov, D., Dashkoff, M., Camacho, S. C., Schumacher, C. A., Irish, J. C., Harkins, T. T., Freeman, E., Garcia, I., Pereira, E., Kendall, S., Belfer, R., Kalir, T., Sebra, R., Reva, B., Dottino, P., and Martignetti, J. A. (2016). "Genomic Analysis of Uterine Lavage Fluid Detects Early Endometrial Cancers and Reveals a Prevalent Landscape of Driver Muta-

tions in Women without Histopathologic Evidence of Cancer: A Prospective Cross-Sectional Study," *PLoS Med* 13(12): e1002206. DOI:10.1371/journal. pmed.1002206.

National Center for Biotechnology Information (2017). https://www.ncbi. nlm.nih.gov/

Niemiec, E., and Howard, H. C. (2016). "Ethical Issues in Consumer Genome Sequencing: Use of Consumers' Samples and Data," *Applied & Translational Genomics* 8: 23–30.

Russek-Cohen, E., Feldblyum, T., Whitaker, K. B., and Hojvat, S. (2011). "FDA Perspectives on Diagnostic Device Clinical Studies for Respiratory Infections," *Clinical Infectious Diseases* 52: S305–S311, DOI: 10.1093/cid/cir056.

Schwartz, L. M., Woloshin, S., Zheng, E., Tse, T., and Zarin, D. A. (2016). "ClinicalTrials.gov and Drugs@FDA: A Comparison of Results Reporting for New Drug Approval Trials," *Annals of Internal Medicine* 165(6): 421–430.

Scientist, The (2016). *Custom Publishing* (October).

Stranneheim, H., and Wedell, A. (2016). "Exome and Genome Sequencing: A Revolution for the Discovery and Diagnosis of Monogenic Disorders," *Journal of Internal Medicine* 279(1): 3–15.

Terluin, B., Eekhout, I., Terwee, C. B., and de Vet, H. C. (2015). "Minimal Important Change (MIC) Based on a Predictive Modeling Approach Was More Precise than MIC Based on ROC Analysis," *Journal of Clinical Epidemiology* 68(12): 1388–1396.

Van Driest, S. L., Wells, Q. S., Stallings, S., Bush, W. S., Gordon, A., Nickerson, D. A., Kim, J. H., Crosslin, D. R., Jarvik, G. P., Carrell, D. S., Ralston, J. D., Larson E. B., Bielinski, S. J., Olson, J. E., Ye, Z., Kullo, I. J., Abul-Husn, N. S., Scott S. A., Bottinger, E., Almoguera, B., Connolly, J., Chiavacci, R., Hakonarson, H., Rasmussen-Torvik, L. J., Pan, V., Persell, S. D., Smith, M., Chisholm, R. L., Kitchner, T. E., He, M. M., Brilliant, M. H., Wallace, J. R., Doheny, K. F., Shoemaker, M. B., Li, R., Manolio, T. A., Callis, T. E., Macaya, D., Williams, M. S., Carey, D., Kapplinger, J. D., Ackerman, M. J., Ritchie, M. D., Denny, J. C., and Roden, D. M. (2016). "Association of Arrhythmia-related Genetic Variants with Phenotypes Documented in Electronic Medical Records," *JAMA* 315(1): 47–57.

Vanhook, P. (2007). "Cost-Utility Analysis: A Method of Quantifying the Value of Registered Nurses," *The Online Journal of Issues in Nursing* 12(3): 5, DOI: 10.3912/OJIN.Vol12No03Man05

Varga, Z., Sinn, P., Fritzsche, F., von Hochstetter, A., Noske, A., Schraml, P., Tausch, C., Trojan, A., and Moch, H. (2013). "Comparison of EndoPredict and Oncotype DX Test Results in Hormone Receptor Positive Invasive Breast Cancer," *PLoS One* 8(3): e58483. doi:10.1371/journal.pone.0058483

Willis, B. H., and Quigley M. (2011). "Uptake of Newer Methodological Developments and the Deployment of Meta-analysis in Diagnostic Test Research: A Systematic Review," *BMC Medical Research Methodology* 11: 27, DOI: 10.1186/1471-2288-11-27.

Wright, D. R., Wittenberg, E., Swan, J. S., Miksad, R. A., and Prosser, L. A. (2009). "Methods for Measuring Temporary Health States for Cost-utility Analyses," *Pharmacoeconomics* 27(9): 713–723.

Yang, T. C., Gryka, A. A., Aucott, L. S., Duthie, G. G., and Macdonald, H. M. (2017). "Longitudinal Study of Weight, Energy Intake and Physical Activity Change across Two Decades in Older Scottish Women," *Journal of Epidemiology & Community Health*, DOI: 10.1136/jech-2016-207948.

Zarin, D. A., Tse, T., Williams, R. J., and Carr, S. (2016). "Trial Reporting in ClinicalTrials.gov – The Final Rule," *New England Journal of Medicine* 375(20): 1998–2004.

Ziegler, K. M., Flamm, C.R., and Aronson, N. (2005). "The Blue Cross Blue Shield Technology Evaluation Center: How We Evaluate Radiology Technologies," *Journal of the American College of Radiology* 2(1): 33–38.

Zhou, X. H., Obuchowiski, N. A., and McClish, D. K. (2011). *Statistical Methods in Diagnostic Medicine*. 2nd edn. Hoboken, NJ: Wiley.

Zou, K. H., Liu, A., Bandos, A. I., Ohno-Machado, L., and Rockette, H. E. (2011). *Statistical Evaluation of Diagnostic Performance: Topics in ROC Analysis*. New York: Chapman & Hall / CRC Press.

Appendix 1 Bayesian Theorem, Proof, and Examples

Simply, if H and E are two disjoint and exhaustive events with probabilities P(H) and P(E) greater than zero in a sample space, the conditional probabilities of H given E, P(H | E), and E given H, P(E | H), may be stated as follows:

P(H | E) = P(H and E) | P(E) and P(E | H) = P(E and H) | P(H) or,

P(H | E)P(E) = P(H and E) and P(H | E)P(H) = P(E and H). (1)

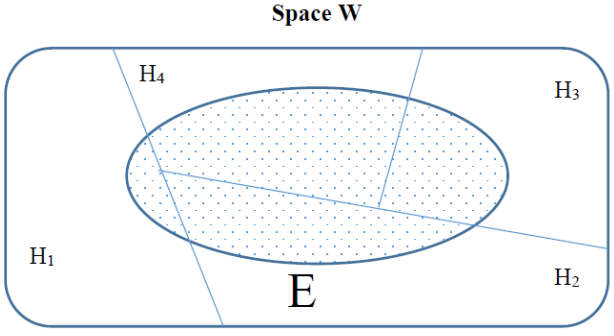
Since the right sides of the equations in (1) are equal, the left sides may be set equal; hence, P(H | E)P(E) = P(E | H)P(H).

Therefore, P(H | E) = P(E | H)P(H) | P(E) (2)

and P(E | H) = P(H | E)P(E) | P(H) (3)

Results (2) and (3) are Bayesian probabilities.

Generally, the Bayes' Theorem and its proof may be stated as follows: Consider the following figure showing intersections in space W of E with events H_1, …, H_4.



As per graph above, let the events H_1, …, H_k form a partition of the space W such that $P(H_j) > 0$ for j = 1, …, k, and let E be any event such that P(E) > 0. Then for i = 1, …, k,

$$P(H_i \mid E) = \frac{P(H_i)P(E \mid H_i)}{\sum_{j=1}^{k} P(H_j)P(E \mid H_j)}. \quad (4)$$

Proof: by the definition of conditional probability, $P(H_i \mid E) = P(H_iE) \mid P(E)$. The numerator on the right side of (4) is equal to $P(H_iE)$ and the denominator is equal to P(E).

To attempt an explanation about the practical usefulness of Bayesian analysis we proceed, without loss of generality, by assuming that theories may be stated, and experiments contacted, in terms of two variables summarized in two-variable contingency tables. Figure A1 portrays a state of nature variable vs. a hypothesis variable.

Figure A1 State of Nature vs. Hypothesis

State of Nature

Hypothesis	**D**	**ND**	**RMT**
+	**a (T+)**	**c (F+)**	**a+c**
–	**b (F-)**	**d (T-)**	**b+d**
CMT	**a+b**	**c+d**	**a+b+c+d**

Where (P=probability):

D = Disease Present
ND = Disease not Present
+ = Test Positive
- = Test Negative
RMT = Row Marginal Total
CMT = Column Marginal Total
Total = N = a+b+c+d
a = True Positive = TP = T+ = Hit
b = False Negative = FN = F**-** = Miss
c = False Positive = FP = F+ = False Alarm
d = True Negative = TN = T- = Correct

P(T+) = a|(a+c) = Positive Predictive Value=PPV
P(F+ given ND) = c|(c+d)
P(F+ given "+") = c|(a+c) = P("+" wrong)
P(T-) = d|(b+d) = Negative Predictive Value = NPV
P(F- given D) = b|(a+b)
P(F- given "-") = b|(b+d) = P("-" wrong)
Prevalence = (a+b)|Total
Sensitivity = a|(a+b)
Specificity = d|(c+d)
P(Overall Accuracy) = (a+d)|Total

Among other tests, the Chi-Square statistic (χ^2_{st}) may be used, when expected frequencies are greater than 5, to examine whether the zero hypothesis H_0 (the state of nature variable and the hypothesis variable are independent) is rejected in favor of the alternative hypothesis H_1 (the state of nature variable and the hypothesis variable are not independent). Expressing the above table as follows,

	D	**ND**	**RMT**
+	A_{ij}	…	TR_1
–	…	…	…
CMT	TC_1	…	**N**

the χ^2_{st}, with degrees of freedom (k−1)(m−1) (k = row and m = column), is

$$\chi^2_{st} = \sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(A_{ij} - E_{ij})^2}{E_{ij}}, \quad \text{where} \quad E_{ij} = \frac{(TR_1)(TC_1)}{N};$$

E_{ij} = expected frequency in the i-th row j-th column.

The H_0 is rejected, in favor of the alternative, when the P-Value is less than 0.05.

Clinical Illustration: Blood Sugar vs. Physical exercise

Let variable one be *blood sugar level* and variable two *physical exercise*. A doctor tests 100 patients for blood sugar levels; 53 of them were diagnosed as high (H) blood sugar patients and 47 as low (L) blood sugar patients. Based on this initial information, the doctor estimates that the probabilities of high and low are, respectively, P(H) = 0.53 and P(L) = 0.47. These probabilities are called the *prior probabilities*.

The doctor, in turn, hypothesizes that frequent exercise contributes to low blood sugar levels; subject to consent and properly designed incentives (e.g., financial or other rewards), she convinces each one of the100 patients to participate in a year-long clinical experiment designed to reveal whether they exercise frequently (Fr) or infrequently (In) subject to a reasonable and objectively chosen cutoff point. The cutoff point may consist of hours of exercise per day / week, or other, above (below) which exercise is classified as frequent (infrequent). At the end of the year-long period, out of the 53 diagnosed as high, 37 exercised infrequently and 16 frequently. Out of the 47 diagnosed as low, 19 exercised infrequently and 28 frequently. The probabilities that emerge from the doctor's experiment are called the *posterior probabilities*. Prior and posterior data as well as statistical tests are summarized in Figure A2. The results indicate that blood sugar levels and exercise are not independent.

The posterior probabilities are reported in Figure A2; they are all Bayesian, computed similarly to P(T+) and P(T-) as shown below:

P(T+) = P(H | In) = [P(In | H)P(H) | P(In)] = [(37 | 53)(53 | 100) | 56 | 100] = 37 | 56 = 0.6607

P(T-) = P(L | Fr) = [P(Fr | L)P(L) | P(Fr)] = [(28 | 47)(47 | 100) | 44 | 100] = 28 | 44 = 0.6364

Figure A2 Blood Sugar vs. Exercise

State of Nature

Exercise	H	L	RMT
In	**37**	**19**	**56**
Fr	**16**	**28**	**44**
CMT	**53**	**47**	**100**

Where (P=probability):

H = High blood sugar
L = Low blood sugar
In = Infrequent exercise
Fr = Frequent exercise
RMT = Row Marginal Total
CMT = Column Marginal Total
Total = N = 100
37 = True Positive = TP = T+ = Hit
19 = False Positive = FP = F+ = False Alarm
16 = False Negative = FN = F- = Miss

28 = True Negative = TN = T- = Correct non-event
P(T+) = 37|56 = 0.6607 = Positive Predictive Value = PPV
P(F+) = 19|56 = 0.3393
P(T-) = 28|44 = 0.6364 = Negative Predictive Value = NPV
P(F-) = 16|44 = 0.3636
Prevalence = 53|100 = 0.53
Sensitivity = 37|53 = 0.6981
Specificity = 28|47 = 0.5957
P(Overall Accuracy) = 65|100 = 0.65

H_0: Blood sugar levels and exercise are independent.
H_1: Blood sugar levels and exercise are not independent.

With degrees of freedom $(2-1)(2-1) = 1$ and $\alpha = 0.05$, the $\chi^2_{Critical} = 3.84$.
Since the $\chi^2_{st} = 8.7299 > \chi^2_{Critical}$, the H_0 is rejected in favor of the alternative.

The table below, produced with STATA, reports the lowest possible P-Value for the χ^2 (also known as Pearson χ^2), the P-Value associated with the Fisher Test (useful for any number of subjects in an experiment) and the expected values below the frequencies.

Exercise	HighLow H	L	Total
AIn	37 29.7	19 26.3	56 56.0
BFr	16 23.3	28 20.7	44 44.0
Total	53 53.0	47 47.0	100 100.0

Pearson chi2(1) = 8.7299 Pr = 0.003
Fisher's exact = 0.005
1-sided Fisher's exact = 0.003

Appendix 2

ID	SUV	Cancer 1=present 2=not present		ID	SUV	Cancer 1=present 2=not present		ID	SUV	Cancer 1=present 2=not present
1	0	2		41	3	2		81	13	1
2	0	2		42	3	2		82	13	2
3	0	2		43	3	2		83	14	1
4	0	2		44	3	2		84	14	2
5	0	2		45	3	2		85	14	1
6	0	2		46	3	2		86	14	1
7	0	2		47	4	1		87	14	1
8	0	2		48	4	2		88	14	2
9	0	2		49	4	2		89	15	1
10	1	2		50	4	1		90	15	1
11	1	2		51	4	2		91	15	1
12	1	2		52	4	2		92	15	1
13	1	2		53	5	1		93	16	1
14	1	2		54	5	2		94	16	1
15	1	2		55	5	2		95	17	1
16	1	2		56	5	2		96	17	1
17	1	2		57	6	1		97	18	1
18	2	2		58	6	2		98	18	1
19	2	2		59	6	1		99	19	1
20	2	2		60	6	2		100	20	1
21	2	2		61	6	2				
22	2	2		62	6	2				
23	2	2		63	7	2				
24	2	2		64	8	1				
25	2	2		65	8	1				
26	2	2		66	9	1				
27	2	2		67	10	2				
28	2	2		68	10	1				
29	2	2		69	11	1				
30	2	2		70	11	1				
31	2	2		71	12	2				
32	2	2		72	12	2				
33	2	2		73	12	1				
34	2	2		74	12	1				
35	2	2		75	13	2				
36	2	2		76	13	1				
37	2	2		77	13	1				
38	2	2		78	13	2				
39	3	2		79	13	1				
40	3	2		80	13	1				