Hindawi
Computational and Mathematical Methods in Medicine
Volume 2019, Article ID 1891569, 22 pages
https://doi.org/10.1155/2019/1891569

Review Article

Medical Diagnostic Tests: A Review of Test Anatomy, Phases, and Statistical Treatment of Data

Sorana D. Bolboacă

Department of Medical Informatics and Biostatistics, Iuliu Haţieganu University of Medicine and Pharmacy Cluj-Napoca, Louis Pasteur Str., No. 6, 400349 Cluj-Napoca, Romania

Correspondence should be addressed to Sorana D. Bolboacă; sbolboaca@gmail.com

Received 30 September 2018; Revised 25 April 2019; Accepted 8 May 2019; Published 28 May 2019

Academic Editor: Juan Pablo Martínez

Diagnostic tests are approaches used in clinical practice to identify with high accuracy the disease of a particular patient and thus to provide early and proper treatment. Reporting high-quality results of diagnostic tests, for both basic and advanced methods, is solely the responsibility of the authors. Despite the existence of recommendation and standards regarding the content or format of statistical aspects, the quality of what and how the statistic is reported when a diagnostic test is assessed varied from excellent to very poor. This article briefly reviews the steps in the evaluation of a diagnostic test from the anatomy, to the role in clinical practice, and to the statistical methods used to show their performances. The statistical approaches are linked with the phase, clinical question, and objective and are accompanied by examples. More details are provided for phase I and II studies while the statistical treatment of phase III and IV is just briefly presented. Several free online resources useful in the calculation of some statistics are also given.

1. Introduction

An accurate and timely diagnostic with the smallest probability of misdiagnosis, missed diagnosis, or delayed diagnosis is crucial in the management of any disease [1, 2]. The diagnostic is an evolving process since both disease (the likelihood and the severity of the disease) and diagnostic approaches evolve [3]. In clinical practice, it is essential to correctly identify the diagnostic test that is useful to a specific patient with a specific condition [4–6]. The over- or underdiagnostic closely reflects on unnecessary or no treatment and harms both the subjects and the health-care systems [3].

Statistical methods used to assess a sign or a symptom in medicine depend on the phase of the study and are directly related to the research question and the design of the experiment (Table 1) [7].

A significant effort was made to develop the standards in reporting clinical studies, both for primary (e.g., case-control studies, cohort studies, and clinical trials) and secondary (e.g., systematic review and meta-analysis) research. The effort led to the publication of four hundred twelve guidelines available on the EQUATOR Network on April 20, 2019 [8]. Each guideline is accompanied by a short checklist describing the information needed to be present in each section and also include some requirements on the presentation of statistical results (information about what, e.g., mean (SD) where SD is the standard deviation, and how to report, e.g., the number of decimals). These guidelines are also used as support in the critical evaluation of an article in evidence-based clinical practice. However, insufficient attention has been granted to the minimum set of items or methods and their quality in reporting the results. Different designs of experiments received more attention, and several statistical guidelines, especially for clinical trials, were developed to standardize the content of the statistical analysis plan [9], for phase III clinical trials in myeloid leukemia [10], pharmaceutical industry-sponsored clinical trials [11], subgroup analysis [12], or graphics and statistics for cardiology [13]. The SAMPL Guidelines provide general

Table 1: Anatomy on phases of a diagnostic test.

Phase	What?	Design
I	Determination of normal ranges (pharmacokinetics, pharmacodynamics, and safe doses)	Observational studies on healthy subjects
II	Evaluation of diagnosis accuracy	Case-control studies on healthy subjects and subjects with the known (by a *gold standard* test) and suspected disease of interest (i) Phase IIa: healthy subjects and subjects with the known disease of interest, all diagnosed by a *gold standard* method (ii) Phase IIb: testing the relevance of the disease severity (evaluate how a test works in ideal conditions) (iii) Phase IIc: assess the predictive values among subjects with suspected disease
III	Evaluation of clinical consequences (benefic and harmful effects) of introducing a diagnostic test	Randomized control trials, randomization determine whether a subject receive or not the diagnosis test
IV	Determination of the long-term consequences of introducing a new diagnostic test into clinical practice	Cohort studies of consecutive participants to evaluate if the diagnostic accuracy of a test in practice corresponds to predictions from systematic reviews of phase III trials

Adapted from [7].

principles for reporting statistical methods and results [14]. SAMPL recommends to provide numbers with the appropriate degree of precision, the sample size, numerator and denominator for percentages, and mean (SD) (where SD = standard deviation) for data approximately normally distributed; otherwise medians and interpercentile ranges, verification of the assumption of statistical tests, name of the test and the tailed (one- or two-tailed), significance level (α), P values even statistically significant or not, adjustment(s) (if any) for multivariate analysis, statistical package used in the analysis, missing data, regression equation with regression coefficients for each explanatory variable, associated confidence intervals and P values, and models' goodness of fit (coefficient of determination) [14]. In regard to diagnostic tests, standards are available for reporting accuracy (QUADAS [15], QUADAS-2 [16], STARD [17, 18], and STARD 2015 [19]), diagnostic predictive models (TRIPOD [20]), systematic reviews and meta-analysis (AMSTAR [21] and AMSTAR 2 [22]), and recommendations and guidelines (AGREE [23], AGREE II [24], and RIGHT [25]). The requirements highlight what and how to report (by examples), with an emphasis on the design of experiment which is mandatory to assure the validity and reliability of the reported results. Several studies have been conducted to evaluate if the available standards in reporting results are followed. The number of articles that adequately report the accuracy is reported from low [26–28] to satisfactory [29], but not excellent, still leaving much room for improvements [30–32].

The diagnostic tests are frequently reported in the scientific literature, and the clinicians must know how a good report looks like to apply just the higher-quality information collected from the scientific literature to decision related to a particular patient. This review aimed to present the most frequent statistical methods used in the evaluation of a diagnostic test by linking the statistical treatment of data with the phase of the evaluation and clinical questions.

2. Anatomy of a Diagnostic Test

A diagnostic test could be used in clinical settings for confirmation/exclusion, triage, monitoring, prognosis, or screening (Table 2) [19, 38]. Table 2 presents the role of a diagnostic test, its aim, and a real-life example.

Different statistical methods are used to support the results of a diagnostic test according to the question, phase, and study design. The statistical analysis depends on the test outcome type. Table 3 presents the most common types of diagnostic test outcome and provides some examples.

The result of an excellent diagnostic test must be accurate (the measured value is as closest as possible by the true value) and precise (repeatability and reproducibility of the measurement) [65]. An accurate and precise measurement is the primary characteristic of a valid diagnostic test.

The reference range or reference interval and ranges of normal values determined in healthy persons are also essential to classify a measurement as a positive or negative result and generally refer to continuous measurements. Under the assumption of a normal distribution, the reference value of a diagnostic measurement had a lower reference limit/lower limit of normal (LRL) and an upper reference limit/upper limit of normal (URL) [66–71]. Frequently, the reference interval takes the central 95% of a reference population, but exceptions from this rule are observed (e.g., cTn-cardiac troponins [72] and glucose levels [73] with <5% deviation from reference intervals) [74, 75]. The reference ranges could be different among laboratories [76, 77], genders and/or ages [78], populations [79] (with variations inclusive within the same population [80, 81]), and to physiological conditions (e.g., pregnancy [82], time of sample collection, or posture). Within-subject biological variation is smaller than the between-subject variation, so reference change values could better reflect the changes in measurements for an individual as compared to reference

Table 2: Anatomy of the role of a diagnostic test.

Role	What?	Example (ref.)
Confirmation/exclusion	Confirm (rule-in) or exclude (rule-out) the disease	Brain natriuretic peptide: diagnostic for left ventricular dysfunction [33]
Triage	An initial test that could be rapidly applied and have a small number of false-positive results	Renal Doppler resistive index: hemorrhagic shock in polytrauma patients [34]
Monitoring	A repeated test that allows assessing the efficacy of an intervention	Glycohemoglobin (A1c Hb): overall glycemic control of patients with diabetes [35]
Prognosis	Assessment of an outcome or the disease progression	PET/CT scan in the identification of distant metastasis in cervical and endometrial cancer [36]
Screening	Presence of the disease in apparently asymptomatic persons	Cytology test: screening of cervical uterine cancer [37]

Table 3: Diagnosis test result: type of data.

Data	Example (ref.)
Qualitative dichotomial	Positive/negative or abnormal/normal (i) Endovaginal ultrasound in the diagnosis of normal intrauterine pregnancy [39] (ii) QuantiFERON-TB test for the determination of tubercular infection [40]
Qualitative ordinal	(i) Prostate bed after radiation therapy: definitely normal/probably normal/uncertain/probably abnormal/definitely abnormal [41] (ii) Scores: Apgar score (assessment of infants after delivery): 0 (no activity, pulse absent, floppy grimace, skin blue or pale, and respiration is absent) to 10 (active baby, pulse over 100 bps, prompt response to stimulation, pink skin, and vigorous cry) [42]; Glasgow coma score: eye opening (from 1 = no eye opening to 4 = spontaneously), verbal response (from 1 = none to 5 = patient oriented), and motor response (from 1 = none to 6 = obeys commands) [43]; Alvarado score (the risk of appendicitis) evaluates 6 clinical items and 2 laboratory measurements and had an overall score from 0 (no appendicitis) to 10 ("very probable" appendicitis) [44]; and sonoelastographic scoring systems in evaluation of lymph nodes [45] (iii) Scales: quality-of-life scales (SF-36 [46], EQ-5D [47, 48], VascuQoL [49, 50], and CIVIQ [51]) and pain scale (e.g., 0 (no pain) to 10 (the worst pain)) [52]
Qualitative nominal	(i) Apolipoprotein E gene (ApoE) genotypes: E2/E2, E2/E3, E2/E4, E3/E3, E3/E4, and E4/E4 [53, 54] (ii) SNP (single-nucleotide polymorphism) of IL-6: at position −174 (rs1800795), −572 (rs1800796), −596 (rs1800797), and T15 A (rs13306435) [55]
Quantitative discrete	(i) Number of bacteria in urine or other fluids [56] (ii) Number of contaminated products with different bacteria [57] (iii) Glasgow aneurysm score (= age in years + 17 for shock + 7 for myocardial disease + 10 for cerebrovascular disease + 14 for renal disease) [58]
Quantitative continuous	(i) Biomarkers: chitotriosidase [59], neopterin [60], urinary cotinine [61], and urinary cadmium levels [61] (ii) Measurements: resistivity index [62], ultrasound thickness [63], and interventricular septal thickness [64]

ranges [83]. Furthermore, a call for establishing the clinical decision limits (CDLs) with the involvement of laboratory professionals had also been emphasized [84].

The Z-score (standardized value, standardized score, or Z-value, Z-score = (measurement $- \mu)/\sigma$)) is a dimensionless metric used to evaluate how many standard deviations (σ) a measurement is far from the population mean (μ) [85]. A Z-score of 3 refers to 3 standard deviations that would mean that more than 99% of the population was covered by the Z-score [86]. The Z-score is properly used under the assumption of normal distribution and when the parameters of the population are known [87]. It has the advantage that allows comparing different methods of measurements [87]. The Z-scores are used on measurements on pediatric population [88, 89] or fetuses [90], but not exclusively (e.g., bone density tests [91]).

3. Diagnostic Tests and Statistical Methods

The usefulness of a diagnostic test is directly related with its reproducibility (the result is the same when two different medical staff apply the test), accuracy (the same result is obtained if the diagnostic test is used more than once), feasibility (the diagnostic method is accessible and affordable), and the effect of the diagnostic test result on the clinical decision [92]. Specific statistical methods are used to sustain the utility of a diagnostic test, and several examples linking the phase of a diagnostic test with clinical question, design, and statistical analysis methods are provided in Table 4 [101].

3.1. Descriptive Metrics. A cohort cross-sectional study is frequently used to establish the normal range of values. Whenever data follow the normal distribution (normality

Table 4: Statistical methods in the assessment of the utility of a diagnostic test.

Phase	Clinical question	Objective(s)	Statistics for results	Example (ref.)
I	Which are the normal ranges of values of a diagnostic test?	Determination of the range of values on healthy subjects	Centrality and dispersion (descriptive) metrics: (i) mean (SD), where SD = standard deviation, if data follow the normal distribution; (ii) otherwise, median ($Q1 - Q3$), where $Q1 = 25^{th}$ percentile and $Q3 = 75^{th}$ percentiles	(i) Levels of hepcidin and prohepcidin in healthy subjects [93] (ii) plasma pro-gastrin-releasing peptide (ProGRP) levels in healthy adults [94]
I	Is the test reproducible?	Variability: (i) Intra- and interobserver (ii) Intra- and interlaboratory	(i) Agreement analysis: % (95% confidence interval) and agreement coefficients (dichotomial data: Cohen, ordinal data: weighted kappa, numerical: Lin's concordance correlation coefficient, and Bland and Altman diagram) (ii) Variability analysis: Coefficient of variation, distribution of differences	(i) Intra- and interobserver variability of uterine measurements [95] (ii) Interlaboratory variability of cervical cytopathology [96] (iii) Concordance between tuberculin skin test and QuantiFERON in children [40]
II	Is the test accurate? Which are performances of the diagnostic test?	Determine the accuracy as compared to a gold standard test	(i) Metrics (dichotomial outcome): Se (sensitivity), Sp (specificity), PPV (predictive positive value), NPV (negative predictive value), and DOR (diagnostic odds ratio) (ii) Clinical performances (dichotomial outcome): PLR (positive likelihood ratio) and NLR (negative likelihood ratio) (iii) Threshold identification (numerical or ordinal with a minimum of five classes outcome): ROC (receiver operating characteristic curve) analysis	(i) Digital breast tomosynthesis for benign and malignant lesions in breasts [97] (ii) Chitotriosidase as a marker of inflammatory status in critical limb ischemia [59] (iii) Sonoelastographic scores to discriminate between benign and malignant cervical lymph nodes [45]
III	Which are the costs, risk, and acceptability of a diagnostic test?	(i) Evaluation of beneficial and harmful effects (ii) Cost-effective analysis	Retrospective or prospective studies: (i) beneficial (e.g., improvement of clinical outcome) or harmful effects (e.g., morbidity and mortality) by proportions, risk ratio, odds ratio, hazard ratio, the number needed to treat, and rates and ratios of desirable or undesirable outcomes (ii) cost-effective analysis (mean cost and quality-adjusted life years (QALYs))	(i) The computed tomography in children, the associated radiation exposure, and the risk of cancer [98] (ii) Healthcare benefit and cost-effectiveness of a screening strategy for colorectal cancer [99]
IV	Which are the consequences of introducing a new diagnostic test into clinical practice?	(i) Does the test result affect the clinical decision?	(i) Studies of pre- and posttest clinical decision-making (ii) %: abnormal, of discrepant results, of tests leading to change the clinical decisions (iii) Costs: per abnormal result, decision change	(i) Does the interferon-gamma release assays (IGRAs) change the clinical management of patients with latent tuberculosis infection (LTBI)? [100]

tests such as Shapiro–Wilk [102] or Kolmogorov–Smirnov test [103, 104] provide valid results whenever the sample sizes exceed 29), the mean and standard deviations are reported [105], and the comparison between groups is tested with parametric tests such as Student's t-test (2 groups) or ANOVA test (more than 2 groups). Median and quartiles ($Q1 - Q3$) are expected to be reported, and the comparison is made with nonparametric tests if experimental data did not follow the normal distribution or the sample size is less than 30 [105]. The continuous data are reported with one or two decimals (sufficient to assure the accuracy of the result), while the P values are reported with four decimals even if the significance threshold was or not reached [106].

The norms and good practice are not always seen in the scientific literature while the studies are frequently more complex (e.g., investigation of changes in the values of biomarkers with age or comparison of healthy subjects with subjects with a specific disease). One example is given by Koch and Singer [107], which aimed to determine the range of normal values of the plasma B-type natriuretic peptide (BNP) from infancy to adolescence. One hundred ninety-five healthy subjects, infants, children, and adolescents were evaluated. Even that the values of BNP varied considerably, the results were improper reported as mean (standard deviation) on the investigated subgroups, but correctly compared subgroups using nonparametric tests [107, 108]. Taheri et al. compared the serum levels of hepcidin (a low molecular weight protein role in the iron metabolism) and prohepcidin in hemodialysis patients (44 patients) and healthy subjects (44 subjects) [93]. Taheri et al. reported the values of hepcidin and prohepcidin as a mean and standard deviation, suggesting the normal distribution of data, and compared using nonparametric tests, inducing the absence of normal distribution of experimental data [93]. Furthermore, they correlated these two biomarkers while no reason exists for this analysis since one is derived from the other [93].

Zhang et al. [94] determined the reference values for plasma pro-gastrin-releasing peptide (ProGrP) levels in healthy Han Chinese adults. They tested the distribution of ProGrP, identified that is not normally distributed, and correctly reported the medians, ranges, and 2.5^{th}, 5^{th}, 50^{th}, 95^{th}, and 97.5^{th} percentiles on two subgroups by ages. Spearman's correlation coefficient was correctly used to test the relation between ProGrP and age, but the symbol of this correlation coefficient was r (symbol attributed to Pearson's correlation coefficient) instead of ρ. The differences in the ProGrP among groups were accurately tested with the Mann–Whitney test (two groups) and the Kruskal–Wallis test (more than two groups). The authors reported the age-dependent reference interval on this specific population without significant differences between genders [94].

The influence of the toner particles on seven biomarkers (serum C-reactive protein (CRP), IgE, interleukin (IL-4, IL-6, and IL-8), serum interferon-γ (IFN-γ), and urine 8-hydroxy-2′-deoxyguanosine (8OHdG)) was investigated by Murase et al. [109]. They conducted a prospective cohort study (toner exposed and unexposed) with a five-year follow-up and measured annually the biomarkers. The reference values of the studied biomarkers were correctly reported as median and 27^{th}–75^{th} percentiles as well as the 2.5^{th}–97.5^{th} percentiles (as recommended by the Clinical and Laboratory Standards Institute [108]).

3.2. Variability Analysis. Two different approaches are used whenever variability of quantitative data is tested in phase I studies, both reflecting the repeated measurements (the same or different device or examiner), namely, variation analysis (coefficient of variation, CV) or the agreement analysis (agreement coefficients).

3.2.1. Variation Analysis. Coefficient of variation (CV), also known as relative standard deviation (RSD), is a standardized measure of dispersion used to express the precision (intra-assay (the same sample assayed in duplicate) CV < 10% is considered acceptable; interassay (comparison of results across assay runs) CV < 15% is deemed to be acceptable) of an assay [110–112]. The coefficient of variation was introduced by Karl Pearson in 1896 [113] and could also be used to test the reliability of a method (the smaller the CV values, the higher the reliability is) [114], to compare methods (the smallest CV belongs to the better method) or variables expressed with different units [115]. The CV is defined as the ratio of the standard deviation to the mean expressed as percentage [116] and is correctly calculated on quantitative data measured on the ratio scale [117]. The coefficient of quartile variation/dispersion (CQV/CQD) was introduced as a preferred measure of dispersion when data did not follow the normal distribution [118] and was defined based on the third and first quartile as $(Q3 - Q1)/(Q3 + Q1)^*$ 100 [119]. In a survey analysis, the CQV is used as a measure of convergence in experts' opinions [120].

The confidence interval associated with CV is expected to be reported for providing the readers with sufficient information for a correct interpretation of the reported results, and several online implementations are available (Table 5).

The inference on CVs can be made using specific statistical tests according to the distribution of data. For normal distributions, tests are available to compare two [121] or more than two CVs (Feltz and Miller test [122] or Krishnamoorthy and Lee test [123], the last one also implemented in R [124]).

Reporting the CVs with associated 95% confidence intervals allows a proper interpretation of its point estimator value (CV). Schafer et al. [125] investigated laboratory reproducibility of urine N-telopeptide (NTX) and serum bone-specific alkaline phosphatase (BAP) measurements with six labs over eight months and correctly reported the CVs with associated 95% confidence intervals. Furthermore, they also compared the CVs between two assays and between labs and highlighted the need for improvements in the analytical precision of both NTX and BAP biomarkers [125]. They concluded with the importance of the availability of laboratory performance reports to clinicians and institutions along with the need for proficiency testing and standardized guidelines to improve market reproducibility [125].

Table 5: Online resources for confidence intervals calculation: coefficient of variation.

What?	URL (accessed on August 26, 2018)
Two-sided confidence interval (CI) for s CV[a]	https://www1.fpl.fs.fed.us/covnorm.dcd.html https://community.jmp.com/kvoqx44227/attachments/kvoqx44227/scripts/77/1/CI%20for%20CV%202.jsl
One-sided CI[a]	
Lower bound	https://www1.fpl.fs.fed.us/covlow.html
Upper bound	https://www1.fpl.fs.fed.us/covup.html
Two-sided CI for s CV[b]	https://www1.fpl.fs.fed.us/covln.html
Ratio of two CVs[a]	https://www1.fpl.fs.fed.us/covratio.html

[a]Normal distribution and [b]lognormal distribution.

However, good practice in reporting CVs is not always observed. Inter- and intra-assay CVs within laboratories reported by Calvi et al. [126] on measurements of cortisol in saliva are reported as point estimators, and neither confidence intervals nor statistical test is provided. Reed et al. [127] reported the variability of measurements (thirty-three laboratories with fifteen repeated measurements on each lab) of human serum antibodies against *Bordetella pertussis* antigens by ELISA method using just the CVs (no associated 95% confidence intervals) in relation with the expected fraction of pairs of those measurements that differ by at least a given factor (k).

3.2.2. Agreement Analysis. Percentage agreement (p_o), the number of agreements divided into the number of cases, is the easiest agreement coefficient that could be calculated but may be misleading. Several agreement coefficients that adjust the proportional agreement by the agreement expected by chance were introduced:

(i) Nominal or ordinal scale: Cohen's kappa coefficient (nominal scale, inclusive dichotomial such as positive/negative test result), symbol κ [128], and its derivatives (Fleiss' generalized kappa [129], Conger's generalized kappa [130], and weighted kappa (ordinal scale test result)) [131]

(ii) Numerical scale: intraclass (Pearson's correlation coefficient (r)) [132] and interclass correlation coefficient (ICC) [133] (Lin's concordance correlation coefficient (ρ_c) [134, 135] and Bland and Altman diagram (B&A plot [136, 137]))

The Cohen's kappa coefficient has three assumptions: (i) the units are independent, (ii) the categories on the nominal scale are independent and mutually exclusive, and (iii) the readers/raters are independent [128]. Cohen's kappa coefficient takes a value between −1 (perfect disagreement) and 1 (complete agreement). The empirical rules used to interpret the Cohen's kappa coefficient [138] are as follows: no agreement for $\kappa \leq 0.20$, minimal agreement for $0.21 < \kappa \leq 0.39$, week agreement for $0.40 \leq \kappa \leq 0.59$, moderate agreement for $0.60 \leq \kappa \leq 0.79$, strong agreement for $0.80 \leq \kappa \leq 0.90$, and almost perfect agreement for $\kappa > 0.90$. The minimum acceptable interrater agreement for clinical laboratory measurements is 0.80. The 95% CI must accompany the value of κ for a proper interpretation, and the empirical interpretation rules must apply to the lower bound of the confidence interval.

The significance of κ could also be calculated, but in many cases, it is implemented to test if the value of κ is significantly different by zero (H_0 *(null hypothesis):* $\kappa = 0$). The clinical significance value is 0.80, and a test using the null hypothesis as H_0: $\kappa = 0.79$ vs. H_1*(one-sided alternative hypothesis):* $\kappa > 0.79$ should be applied.

Weighted kappa is used to discriminate between different readings on ordinal diagnostic test results (different grade of disagreement exists between *good* and *excellent* compared to *poor* and *excellent*). Different weights reflecting the importance of agreement and the weights (linear, proportional to the number of categories apart or quadratic, proportional to the square of the number of classes apart) must be established by the researcher [131].

Intra- and interclass correlation coefficients (ICCs) are used as a measure of reliability of measurements and had their utility in the evaluation of a diagnostic test. Interrater reliability (defined as two or more raters who measure the same group of individuals), test-retest reliability (defined as the variation in measurements by the same instrument on the same subject by the same conditions), and intrarater reliability (defined as variation of data measured by one rater across two or more trials) are common used [139]. McGraw and Wong [140] defined in 1996 the ten forms of ICC based on the *model* (1-way random effects, 2-way random effects, or 2-way fixed effects), the *number of rates/measurements* (single rater/measurement or the mean of k raters/measurements), and *hypothesis* (consistency or absolute agreement). McGraw and Wong also discuss how to correctly select the correct ICC and recommend to report the ICC values along with their 95% CI [140].

Lin's concordance correlation coefficient (ρ_c) measures the concordance between two observations, one measurement as the *gold standard*. The ranges of values of Lin's concordance correlation coefficient are the same as for Cohen's kappa coefficient. The interpretation of ρ_c takes into account the scale of measurements, with more strictness for continuous measurements (Table 6) [141, 142]. For intra- and interobserver agreement, Martins and Nastri [142] introduced the metric called limits of agreement (LoA) and proposed a cutoff < 5% for very good reliability/agreement.

Reporting the ICC and/or CCC along with associated 95% confidence intervals is good practice for agreement

Table 6: Intra- and interclass correlation coefficients and concordance correlation coefficient: an empirical assessment of the strength of agreement.

Agreement	Continuous measurement	Ultrasound fetal measurements	Semiautomated measurements
Very good	$\rho_c > 0.99$	$\rho_c > 0.998$	$\rho_c > 0.90$
Good	$0.95 < \rho_c \leq 0.99$	$0.99 < \rho_c \leq 0.998$	$0.80 < \rho_c \leq 0.90$
Moderate	$0.90 < \rho_c \leq 0.95$	$0.98 < \rho_c \leq 0.99$	$0.65\rho_c \leq 0.80$
Poor	$0.70 < \rho_c \leq 0.90$	$0.95 < \rho_c \leq 0.98$	$\rho_c < 0.65$
Very poor	$\rho_c < 0.70$	$\rho_c < 0.95$	

Source [141, 142].

analysis. The results are reported in both primary (such as reliability analysis of the Microbleed Anatomical Rating Scale in the evaluation of microbleeds [143], automatic analysis of relaxation parameters of the upper esophageal sphincter [144], and the use of signal intensity weighted centroid in magnetic resonance images of patients with discs degeneration [145]) and secondary research studies (systematic review and/or meta-analysis: evaluation of the functional movement screen [146], evaluation of the Manchester triage scale on an emergency department [147], reliability of the specific physical examination tests for the diagnosis of shoulder pathologies [148], etc.).

Altman and Bland criticized the used of correlation (this is a measure of association, and it is not correct to infer that the two methods can be used interchangeably), linear regression analysis (the method has several assumptions that need to be checked before application, and the assessment of residuals is mandatory for a proper interpretation), and the differences between means as comparison methods aimed to measure the same quantity [136, 149, 150]. They proposed a graphical method called the B&A plot to analyze the agreement between two quantitative measurements by studying the mean difference and constructing limits of agreement [136, 137]. Whenever a *gold standard* method exists, the difference between the two methods is plotted against the reference values [151]. Besides the fact that the B&A plot provides the limits of agreements, no information regarding the acceptability of the boundaries is supplied, and the acceptable limits must be a priori defined based on clinical significance [150]. The B&A plot is informally interpreted in terms of bias (*How big the average discrepancy between the investigated methods is? Is the difference large enough to be clinically relevant?*), equivalence (*How wide are the limits of agreement?*, limits wider than those defined clinically indicate ambiguous results while narrow and small bias suggests that the two methods are equivalent), and trend and variability (*Are the dots homogenous*?).

Implementation of the 95% confidence intervals associated to ICC, CCC, or kappa statistics and the test of significance are implemented in commercial or free access statistical programs (such as SPSS, MedCalc, SAS, STATA, R, and PASS-NCSS) or could be found freely available online (e.g. vassarstats-©Richard Lowry 2001–2018, http://vassarstats.net/kappa.html; KappaCalculator ©Statistics Solutions 2018, http://www.statisticssolutions.com/KappaCalculator.html; and KappaAcc-Bakeman's Programs, http://bakeman.gsucreate.org/kappaacc/; all accessed August 27, 2018)).

3.3. Accuracy Analysis. The accuracy of a diagnostic test is related to the extent that the test gives the right answer, and the evaluations are done relative to the best available test (also known as *gold standard* test or *reference* test and hypothetical ideal test with sensitivity (Se) = 100% and specificity (Sp) = 100%) able to reveal the right answer. Microscopic examinations are considered the *gold standard* in the diagnosis process but could not be applied to any disease (e.g., stable coronary artery disease [152], rheumatologic diseases [153], psychiatric disorders [154], and rare diseases with not yet fully developed histological assessment [155]).

The factors that could affect the accuracy of the diagnostic test can be summarized as follows [156, 157]: sampling bias, imperfect *gold standard* test, artefactual variability (e.g., changes in prevalence due to inappropriate design) or clinical variability (e.g., patient spectrum and "gold-standard" threshold), subgroups differences, or reader expectations.

Several metrics calculated based on the 2×2 contingency table are frequently used to assess the accuracy of a diagnostic test. A *gold standard* or *reference* test is used to classify the subject either in the group with the disease or in the group without the disease of interest. Whatever the type of data for the diagnostic test is, a 2×2 contingency table can be created and used to compute the accuracy metrics. The generic structure of a 2×2 contingency table is presented in Table 7, and if the diagnostic test is with high accuracy, a significant association with the reference test is observed (significant Chi-square test or equivalent (for details, see [158])).

Several standard indicators and three additional metrics useful in the assessment of the accuracy of a diagnostic test are briefly presented in Tables 8 and 9.

The reflection of a positive or negative diagnosis on the probability that a patient has/not a particular disease could be investigated using Fagan's diagram [165]. The Fagan's nomogram is frequently referring in the context of evidence-based medicine, reflecting the decision-making for a particular patient [166]. The Bayes' theorem nomogram was published in 2011, the method incorporating in the prediction of the posttest probability the following metrics: pretest probability, pretest odds (for and against), PLR or NLR, posttest odds (for and against), and posttest probability [167]. The latest form of Fagan's nomogram, called two-step Fagan's nomogram, considered pretest probability, Se (Se of test for PLR), LRs, and Sp (Sp of test for NLR), in predicting the posttest probability [166].

Table 7: 2×2 contingency generic table.

Diagnostic test result	Disease present	Disease absent	Total
Positive	TP (true positive)	FP (false positive)	TP + FP
Negative	FN (false negative)	TN (true negative)	FN + TN
Total	TP + FN	FP + TN	n = TP + FP + FN + TN

Total on the rows represents the number of subjects with positive and respectively negative test results; total on the columns represents the number of subjects with (disease present) and respectively without (disease absent) the disease of interest; and the classification as test positive/test negative is done using the cutoff value for ordinal and continuous data.

Table 8: Standard statistic indicators used to evaluate diagnostic accuracy.

Statistic (Abb)	Formula	Remarks
Sensitivity (Se)	TP/(TP + FN)	(i) The highest the Se, the smallest the number of false negative results (ii) High Se: (a) a negative result rules-out (SnNOUT) (b) suitable for screening (ruling-out)
Specificity (Sp)	TN/(TN + FP)	(i) The highest the Se, the smallest the number of false-positive results (ii) High Sp: (a) a positive result rules-in (SpPIN) (b) It is suitable for diagnosis (ruling-in)
Accuracy index (AI)	(TP + TN)/(TP + FP + FN + TN)	(i) Give information regarding the cases with the right diagnosis (ii) It is difficult to convert its value to a tangible clinical concept (iii) It is affected by the prevalence of the disease
Youden's index (J) [159]	Se + Sp − 1	(i) Sums the cases wrongly classified by the diagnostic test (ii) Assess the overall performance of the test. $J = 0$, if the proportion of positive tests is the same in the group with/without the disease. $J = 1$, if no FPs or FNs exist (iii) Misleading interpretation in comparison of the effectiveness of two tests (iv) Used to identify the best cutoff on ROC analysis: its maximum value corresponds to the highest distance from diagonal
Positive predictive value (PPV)*	TP/(TP + FP)	(i) Answer the question "what is the chance that a person with a positive test truly has the disease?" (ii) Clinical applicability for a particular subject with a positive test result (iii) It is affected by the prevalence of the disease
Negative predictive value (NPV)*	TN/(TN + FN)	(i) Answer the question "what is the chance that a person with a negative test truly not to have the disease?" (ii) Clinical applicability for a particular subject with a negative test result (iii) It is affected by the prevalence of the disease
Positive likelihood ratio (PLR/LR+)*	Se/(1 − Sp)	(i) Indicates how much the odds of the disease increase when a test is positive (indicator to rule-in) (ii) PLR (the higher, the better) (a) $> 10 \longrightarrow$ convincing diagnostic evidence (b) $5 < \text{PLR} < 10 \longrightarrow$ strong diagnostic evidence
Negative likelihood ratio (NLR/LR−)*	(1 − Se)/Sp	(i) Indicates how much the odds of the disease decrease when a test is negative (indicator to rule-out) (ii) NLR (the lower, the better) (a) $< 0.1 \longrightarrow$ convincing diagnostic evidence (b) $0.2 < \text{PLR} < 0.1 \longrightarrow$ strong diagnostic evidence

Table 8: Continued.

Statistic (Abb)	Formula	Remarks
Diagnostic odds ratio (DOR)** [160]	(TP/FN)/(FP/TN) [Se/(1 – Se)]/[(1 – Sp)/Sp] [PPV/(1 – PPV)]/[(1 – NPV)/NPV] PLR/NLR	(i) High DOR indicates a better diagnostic test performance (ranges from 0 to infinite). A value of 1 indicates a test not able to discriminate between those with and those without the disease (ii) Combines the strengths of Se and Sp (iii) Useful to compare different diagnostic tests (iv) Not so useful when the aim is to *rules-in* or *rules-out* (v) Convenient indicator in the meta-analysis
Posttest odds (PTO)* Posttest probability (PTP)*	Pretest odds (prevalence/(1 – prevalence)) × LR PTO/(PTO + 1)	(i) Gives the odds that the patient has to the target disorder after the test is carried out (ii) Gives the proportion of patients with that particular test result who have the target disorder

All indicators excepting J are reported with associated 95% confidence intervals; ROC = receiver-operating characteristic; *patient-centered indicator; TP = true positive; FP = false positive; FN = false negative; TN = true negative; and PPV and NPV depend on the prevalence (to be used only if (no. of subjects with disease)/(no. of patients without disease) is equivalent with the prevalence of the disease in the studied population).

Table 9: Other metrics used to evaluate diagnosis accuracy.

Statistic (Abb)	Formula	Remarks
Number needed to diagnose (NND) [161]	$1/[\text{Se} - (1 - \text{Sp})]1/J$	(i) The number of patients that need to be tested to give one correct positive test result (ii) Used to compare the costs of different tests
Number needed to misdiagnose (NNM) [162]	$1/[1 - (\text{TP} + \text{TN})/n]$	(i) The highest the NNM, the better the diagnostic test
Clinical utility index (CUI) [163, 164]	$\text{CUI+} = \text{Se} \times \text{PPV}$ $\text{CUI−} = \text{Sp} \times \text{NPV}$	(i) Gives the degree to which a diagnostic test is useful in clinical practice (ii) Interpretation: $\text{CUI} > 0.81 \longrightarrow$ *excellent utility*; $0.64 \leq \text{CUI} < 0.81 \longrightarrow$ *good utility*; $0.49 \leq \text{CUI} < 0.64 \longrightarrow$ *fair utility*; $0.36 \leq \text{CUI} < 0.49 \longrightarrow$ *poor utility*; and $\text{CUI} < 0.36 \longrightarrow$ *very poor utility*

Abb = abbreviation; all indicators excepting J are reported with associated 95% confidence intervals; TP = true positive; FP = false positive; FN = false negative; and TN = true negative.

The receiver operating characteristic (ROC) analysis is conducted to investigate the accuracy of a diagnostic test when the outcome is quantitative or ordinal with at least five classes [168, 169]. ROC analysis evaluates the ability of a diagnostic test to discriminate positive from negative cases. Several metrics are reported related to the ROC analysis in the evaluation of a diagnostic test, and the most frequently used metrics are described in Table 10 [170, 171]. The closest the left-upper corner of the graph, the better the test. Different metrics are used to choose the cutoff for the optimum Se and Sp, such as Youden's index (J, maximum), d^2 ($(1 - \text{Se})^2 + (1 - \text{Sp})^2$, minimum), the weighted number needed to misdiagnose (maximum, considered the pretest probability and the cost of a misdiagnosis) [172], and Euclidean index [173]. The metrics used to identify the best cutoff value are a matter of methodology and are not expected to be reported as a result (reporting a J index of 0.670 for discrimination in small invasive lobular carcinoma [174] is not informative because the same J could be obtained for different values of Se and Sp: 0.97/0.77, 0.7/0.97, 0.83/0.84, etc.). Youden's index has been reported as the best metric in choosing the cutoff value [173] but is not able to differentiate between differences in sensitivity and specificity [175]. Furthermore, Youden's index can be used as an indicator of quality when reported with associated 95% confidence intervals, and a poor quality being associated with the presence of 0.5 is the confidence interval [175].

3.4. Performances of a Diagnostic Test by Examples. The body mass index (BMI) was identified as a predictor marker of breast cancer risk on Iranian population [176], with an AUC 0.79 (95% CI: 0.74 to 0.84).

A simulation dataset was used to illustrate how the performances of a diagnostic test could be evaluated, evaluating the BMI as a marker for breast cancer. The simulation was done with respect to the normal distribution for 100 cases with malign breast tumor and 200 cases with benign breast tumors with BMI mean difference of 5.7 kg/m^2 (Student's t-test assuming unequal variance: t-stat = 9.98, $p < 0.001$). The body mass index (BMI) expressed in kg/m^2 varied from 20 to 44 kg/m^2, and the ROC curve with associated AUC is presented in Figure 1.

The ROC curve graphically represents the pairs of Se and (1 – Sp) for different cutoff values. The AUC of 0.825 proved significantly different by 0.5 ($p < 0.001$), and the point estimator indicates a good accuracy, but if the evaluation is done based on the interpretation of the 95% lower bound, we

Table 10: Metrics for global test accuracy evaluation or comparisons of performances of two tests.

Statistic (Abb)	Method	Remarks
Area under the ROC curve (AUC)	(i) Nonparametric (no assumptions): empirical method (estimated AUC is biased if only a few points are in the curve) and smoothed-curve methods such as kernel density method (not reliable near the extremes of the ROC curve) (ii) Parametric (the distributions of the cases and controls are normal): binomial method (tighter asymptotic confidence bounds for samples less than 100)	(i) AUC = 1 ⟶ perfect diagnostic test (perfect accuracy) (ii) AUC ~ 0.5 ⟶ random classification (iii) $0.9 < AUC \leq 1$ ⟶ excellent accuracy classification (iv) $0.8 < AUC \leq 0.9$ ⟶ good accuracy (v) $0.7 < AUC \leq 0.8$ ⟶ worthless
Partial area under the curve (pAUC)	(i) Nonparametric (no assumptions) (ii) Parametric: using the binomial assumption	(i) Looks to a portion AUC for a predefined range of interest (ii) Depends on the scale of possible values on the range of interest (iii) Has less statistical precision compared to AUC
Diagnostic odds ratio (DOR)	(i) Must use the same fixed cutoff (ii) Most useful in a meta-analysis when two or more tests are compared	(i) DOR = 1 ⟶ test (ii) DOR increases as ROC is closer to the top left-hand corner of the ROC plot (iii) The same DOR could be obtained for different combinations of Se and Sp
TP fraction for a given FP fraction (TPF_{FPF})	(i) Need the same false-positive fraction	(i) Useful to compare two different tests at a specific FPF (decided based on clinical reasoning), especially when the ROC curves cross
Comparison of two tests	(i) Comparison of AUC of two different tests (ii) Absolute difference ($Se_A - Se_B$) or ratio (Se_A/Se_B), where A is one diagnostic test and B is another diagnostic test	(i) Apply the proper statistical test; each AUC must be done relative to the "gold-standard" test (ii) Test A better than B if absolute difference is > 0; ratio > 1

Abb = abbreviation; all indicators are reported with associated 95% confidence intervals; *patient-centered indicator; TP = true positive; FP = false positive; FN = false negative; and TN = true negative.

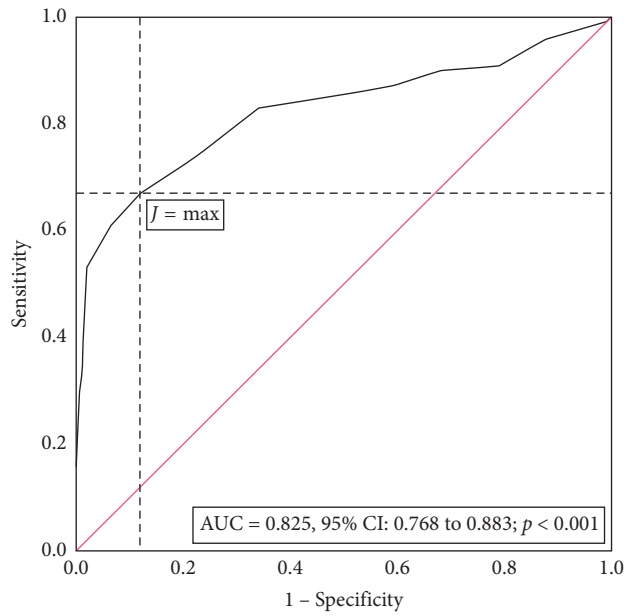

Figure 1: Summary receiver operating characteristic (ROC) curve for BMI as an anthropometric marker to distinguish benign from malign breast tumors. The red line shows an equal proportion of correctly classified breast cancer sample and incorrectly classifies samples without breast cancer (random classification). The J max (max (Se + Sp − 1)) corresponds to a Se = 0.67 and a Sp = 0.88 for a cutoff > 29.5 kg/m^2 (BMI) for the breast cancer sample.

found the BMI as a worthless test for breast cancer. The J had its maximum value at a cutoff equal to 29.5 kg/m^2 and corresponded to a Se of 0.67, a Sp of 0.88, and an AI of 0.81. The PLR of 5.58 indicates that the BMI is strong diagnostic evidence, but this classification is not supported by the value of NLR which exceed the value of 0.2 (Table 10). A BMI >29.5 kg/m^2 *usually occurs* in those with breast cancer while a BMI ≤ 29.5 kg/m^2 *often occurs* in those without breast cancer. At a cutoff of 29.5 kg/m^2, the marker is *very poor* for finding those *with breast cancer* but is good for screening.

The performance metrics varied according to the cutoff values (Table 11). A cutoff with a low value is chosen whenever the aim is to minimize the number of false negatives, assuring a Se of 1 (19.5 kg/m^2, TP = 100, Table 10). If a test able to correctly classify the true negatives is desired, the value of the cutoff must be high (38.5 kg/m^2, TN = 200, Table 11) assuring a Sp of 1.

The analysis of the performance metrics for our simulation dataset showed that the maximum CUI+ and CUI− values are obtained for the cutoff value identified by the J index, supporting the usefulness of the BMI for screening not for case finding.

The accuracy analysis is reported frequently in the scientific literature both in primary and secondary studies. Different actors such as the authors, reviewers, and editors could contribute to the quality of the statistics reported. The evaluation of plasma chitotriosidase as a biomarker in critical

Table 11: Performances metrics for body mass index (BMI) as an anthropometric marker for breast cancer.

Indicator	Cutoff–BMI (kg/m^2)						
	19.5	22.5	25.5	29.5	32.5	35.5	38.5
TP (true positives)	100	96	87	67	43	25	13
FP (false positives)	200	176	117	24	3	1	0
TN (true negatives) off	0	24	83	176	197	199	200
FN (false negatives)	0	4	13	33	57	75	87
Se (sensitivity)	1	1	0.87	0.67	0.43	0.25	0.13
Sp (specificity)	0	0.10	0.42	0.88	0.99	0.99	1
PPV (positive predictive value)	0.33	0.40	0.43	0.74	0.94	0.96	1
NPV (negative predictive value)	n.a.	0.90	0.87	0.84	0.78	0.73	0.70
PLR (positive likelihood ratio)	1.00	1.10	1.49	5.58	28.7	50.0	n.a.
NLR (negative likelihood ratio)	n.a.	0.30	0.31	0.38	0.58	0.75	0.84
AI (accuracy index)	0.33	0.40	0.57	0.81	0.80	0.75	0.71
CUI+ (clinical utility index positive)	0.33	0.30	0.37	0.47	0.40	0.24	0.13
CUI− (clinical utility index negative)	n.a.	10	0.36	0.74	0.76	0.72	0.70

limb ischemia reported the AUC with associated 95% confidence intervals, cutoff values [59], but no information on patient-centered metrics or utility indications are provided. Similar parameters as reported by Ciocan et al. [59] have also been reported in the evaluation of sonoelastographic scores in the differentiation of benign by malign cervical lymph nodes [45]. Lei et al. conducted a secondary study to evaluate the accuracy of the digital breast tomosynthesis versus digital mammography to discriminate between malign and benign breast lesions and correctly reported Se, Sp, PLR, NLR, and DOR for both the studies included in the analysis and the pooled value [97]. However, insufficient details are provided in regard to ROC analysis (e.g., no AUCs confidence intervals are reported) or any utility index [97]. Furthermore, Lei et al. reported the Q* index which reflect the point on the SROC (summary receiver operating characteristic curve) at which the Se is equal with Sp that could be useful in specific clinical situations [97].

The number needed to diagnose (NND) and number needed to misdiagnose (NNM) are currently used in the identification of the cutoff value on continuous diagnostic test results [172, 177], in methodological articles, or teaching materials [161, 178, 179]. The NND and NNM are less frequently reported in the evaluation of the accuracy of a diagnostic test. Several examples identified in the available scientific literature are as follows: color duplex ultrasound in the diagnosis of carotid stenosis [180], culture-based diagnosis of tuberculosis [181], prostate-specific antigen [182, 183], endoscopic ultrasound-guided fine needle biopsy with 19-gauge flexible needle [184], number needed to screen-prostate cancer [185, 186], the integrated positron emission tomography/magnetic resonance imaging (PET/MRI) for segmental detection/localization of prostate cancer [187], serum malondialdehyde in the evaluation of exposure to chromium [188], the performances of the matrix metalloproteinase-7 (MMP-7) in the diagnosis of epithelial injury and of biliary atresia [189], lactate as a diagnostic marker of pleural and abdominal exudate [190], the Gram stain from a joint aspiration in the diagnosis of pediatric septic arthritis [191], and performances of a sepsis algorithm in an emergency department [192]. Unfortunately, the NND or NNM point estimators are not all the time reported with the associated 95% confidence intervals [161, 180, 181, 186, 187, 190, 191].

The reporting of the clinical utility index (CUI) is more frequently seen in the evaluation of a questionnaire. The grades not the values of CUIs were reported by Michell et al. [193] in the assessment of a semistructured diagnostic interview as a diagnostic tool for the major depressive disorder. Johansson et al. [194] reported both the CUI + value and its interpretation in cognitive evaluation using Cognistat. The CUI+/CUI− reported by Michell et al. [195] on the patient health questionnaire for depression in primary care (PHQ-9 and PHQ-2) is reported as a value with associated 95% confidence interval as well as interpretation. The CUI+ and CUI− values and associated confidence intervals were also reported by Fereshtehnejad et al. [196] in the evaluation of the screening questionnaire for Parkinsonism but just for the significant items. Fereshtehnejad et al. [196] also used the values of CUI+ and CUI− to select the optimal screening items whenever the value of point estimator was higher than 0.63. Bartoli et al. [197] represented the values of CUI graphically as column bars (not necessarily correct since the CUI is a single value, and a column could induce that is a range of values) in the evaluation of a questionnaire for alcohol use disorder on different subgroups. The accurate reporting of CUIs as values and associated confidence intervals could also be seen in some articles [198, 199], but is not a common practice [200–207].

Besides the commercial statistical programs able to assist researchers in conducting an accuracy analysis for a diagnostic test, several free online (Table 12) or offline applications exist (CATmaker [208] and CIcalculator [209]).

Smartphone applications have also been developed to assist in daily clinical practice. The *DocNomo* application for iPhone/iPad free application [210] allows calculation of posttest probability using the two-step Fagan nomogram. Other available applications are *Bayes' posttest probability calculator*, *EBM Tools app*, and *EBM Stats Calc*. Allen et al. [211] and Power et al. [212] implemented two online tools for the visual examination of the effect of Se, Sp, and prevalence on TP, FP, FN, and TN values and the evaluation

Table 12: Online applications for diagnostic tests: characteristics.

Name	Input	Output
Diagnostic test calculator[a]	TP, FP, TN, FN *OR* Prevalence AND Se AND Sp AND sample size *OR* Prevalence AND PLR AND NLR AND sample size	Prevalence AND Se AND Sp AND PLR AND NLR Fagan diagram
Diagnostic test calculator evidence-based medicine toolkit[b]	TP, FP, TN, FN	Se, Sp, PPV, NPV, PLR, NLR with associated 95% confidence intervals Posttest probability graph
MedCalc: Bayesian analysis model[c]	Prevalence AND Se AND Sp *OR* TP, FP, TN, FN	PPV, NPV, LPR, NLR, posttest probability
MedCalc[d]	TP, FP, TN, FN	Se, Sp, PPV, NPV, PLR, NLR, prevalence, AI with associated 95% confidence intervals
Clinical calculator 1[e]	TP, FP, TN, FN	Se, Sp, PPV, NPV, PLR, NLR, prevalence, AI with associated 95% confidence intervals
Clinical utility index calculator[f]	TP, TN, total number of cases, the total number of noncases	Se, Sp, PPV, NPV, PLR, NLR, prevalence, AI with associated 95% confidence intervals
DiagnosticTest[g]	Number of positive and negative gold standard results for each level of the new diagnostic test	Se, Sp, PPV, NPV, PLR, NLR, AI, DOR, Cohen's kappa, entropy reduction, and a bias Index ROC curve if > 2 levels for all possible cutoff
Simple ROC curve analysis[h]	Absolute frequencies for false positive and the true positive for up to ten diagnostic levels	Cumulative rates (false positive and true positive) and ROC curve (equation, R^2, and AUC)
ROC analysis[i]	Five different type of input data: an example for each type is provided	Se, Sp, AI, positive cases missed, negative cases missed, AUC, ROC curve
AUSVET: EpiTools[j]	TP, FP, TN, FN	Different tools from basic accuracy to comparison of two diagnostic tests to ROC analysis

All URLs were retrieved on April 20, 2019. TP = true positive; FP = false positive; FN = false negative; TN = true negative; Se = sensitivity; Sp = specificity; AI = accuracy index; PPV = positive predictive value; NPV = negative predictive value; PLR = positive likelihood ratio; NLR = negative likelihood ratio; DOR = diagnostic odds ratio; ROC = receiver operating characteristic; AUC = area under the ROC curve; [a]http://araw.mede.uic.edu/cgi-bin/testcalc.pl; [b]https://ebm-tools.knowledgetranslation.net/calculator/diagnostic/; [c]http://www.medcalc.com/bayes.html; [d]https://www.medcalc.org/calc/diagnostic_test.php; [e]http://vassarstats.net/clin1.html; [f]http://www.psycho-oncology.info/cui.html; [g]http://www.openepi.com/DiagnosticTest/DiagnosticTest.htm; [h]http://vassarstats.net/roc1.html; [i]http://www.rad.jhmi.edu/jeng/javarad/roc/JROCFITi.html; and [j]http://epitools.ausvet.com.au/content.php?page=TestsHome.

of clinical accuracy and utility of a diagnostic test [213]. Furthermore, they have underconstructed the evaluation of the uncertainties in assessing test accuracy when the reference standard is not perfect as support for the evidence-based practice.

4. Cost-Benefit Analysis

The studies conducted in phase III and IV in the investigation of a diagnostic test could be covered under the generic name of cost-benefit analysis. Different aspects of the benefit could be investigated such as societal impact (the impact on the society), cost-effectiveness (affordability), clinical efficacy or effectiveness (effects on the outcome), cost-consequence analysis, cost-utility analysis, sensitivity analysis (probability of disease and/or recurrence, cost of tests, impact on QALY (quality-adjusted life-year), and impact of treatment), and analytical performances (precision, linearity, and cost-effectiveness ratio) [214]. Thus, the evaluation of diagnostic tests benefits could be investigated from different perspectives (e.g., societal, health-care system, and health-care provider) and considering different items (e.g., productivity, patient and family time, medication, and physician time) [215]. Furthermore, an accurate comparison of two diagnostic tests must consider both the accuracy and benefit/harm in the assessment of the clinical utility [216, 217]. Generally, then cost-benefit analysis employs multivariate and multifactorial analysis using different designs of the experiment, including survival analysis, and the statistical approach is selected according to the aim of the study. Analysis of relationships is done using correlation method (Person's correlation (r) when the variables (two) are quantitative and normal distributed, and a linear relation is assuming between them; Spearman's (ρ) or Kendall's (τ) correlation coefficient otherwise; it is recommended to use Kendall's tau instead of Spearman's rho when data have ties [218]) or regression analysis when the nature of the relationship is of interest and an outcome

variable exists [219]. The statistical methods applied when cost-benefit analysis is of interest are not discussed in detail here, but the basic requirements in reporting results are as follows [220–225]:

(i) Correlation analysis: give summary statistic according to the distribution of data (with associated 95% confidence intervals when appropriate, for both baseline data and outcome data), graphical representation as scatter plot, use correct symbol of the correlation coefficient and associate the P value along with the sample size, report missing data, and report the check for influential/outliers.

(ii) Multivariate or multifactorial analysis: summary of the check of assumptions (plots, tests, and indicators), provide the plot of the model, give the model with coefficients, standard error of the coefficients and associated P values or 95% confidence intervals, determination coefficient of the mode, standard error of the model, statistic and P value of the model, provide the sample size, give the number of missing data for each predictor, and adjusted and unadjusted metrics (e.g., OR in logistic regression and HR (hazard ratio) in survival analysis).

Miglioretti et al. [98] investigated the link between radiation exposure of children through the CT examination and the risk of cancer. They reported a trend of the use in the CT which increased from 1996 to 2005, a plateau between 2005 and 2007 followed by a decrease till 2010. The number of CT scans was reported per 1,000 children. Regardless of the anatomical CT scan, the average effective doses were expressed as mean and percentiles (25^{th}, 50^{th}, 75^{th}, and 95^{th}), while the dose exceeding 20 mSv was reported as percentages. The mean organ dose was also reported and the lifetime attributable risk of solid cancer or leukemia, as well as some CT scans leading to one case of cancer per 10,000 scans [98]. The reported numbers and risks were not accompanied by the 95% confidence intervals [98] excepting the estimated value of the total number of future radiation-induced cancers related to pediatric CT use (they named it as uncertainty limit).

Dinh et al. [99] evaluated the effectiveness of a combined screening test (fecal immunological test and colonoscopy) for colorectal cancer using the Archimedes model (human physiology, diseases, interventions, and health-care systems [226]). The reported results, besides frequently used descriptive metrics, are the health utility score [227], cost per person, quality-adjusted life-years (QALYs) gained per person, and cost/QALYs gain as numerical point estimators not accompanied by the 95% confidence interval.

Westwood et al. [228] conducted a secondary study to evaluate the performances of the high-sensitivity cardiac troponin (hs-cTn) assays in ruling-out the patients with acute myocardial infarction (AMI). Clinical effectiveness using metrics such as Se, Sp, NLR, and PLR (for both any threshold and 99^{th} percentile threshold) was reported with associated 95% confidence intervals. As the cost-effectiveness metrics the long-term costs, cost per life-year (LY) gained, quality-adjusted life-years (QALYs), and costs/QALYs were reported with associated 95% confidence intervals for different Tn testing methods. Furthermore, the incremental cost-effectiveness ratio (ICER) was used to compare the mean costs of two Tn testing methods along with the multivariate analysis (reported as estimates, standard error of the estimate, and the distribution of data).

Tiernan et al. [100] reported the changes in the clinical practice for the diagnosis of latent tuberculosis infection (LTBI) with interferon-gamma release assay, namely, QuantiFERON-TB Gold In-Tube (QFT, Cellestis, Australia). Unfortunately, the reported outcome was limited to the number of changes in practice due to QFT as absolute frequency and percentages [100].

5. Limitations and Perspectives

The current paper did not present either detail regarding the research methodology for diagnostic studies nor the critical appraisal of a paper presenting the performances of a diagnostic test because these are beyond the aim. Extensive scientific literature exists regarding both the design of experiments for diagnostic studies [4, 15, 92, 229, 230] and the critical evaluation of a diagnostic paper [231–234]. As a consequence, neither the effect of the sample size on the accuracy parameters, or the *a priori* computation of the sample size needed to reach the level of significance for a specific research question, nor the *a posteriori* calculation of the power of the diagnostic test is discussed. The scientific literature presenting the sample size calculation for diagnostic studies is presented in the scientific literature [235–238], but these approaches must be used with caution because the calculations are sensitive and the input data from one population are not a reliable solution for another population, so the input data for sample size calculation are recommended to come from a pilot study. This paper does not treat how to select a diagnostic test in clinical practice, the topic being treated by the evidence-based medicine and clinical decision [239–241].

Health-care practice is a dynamic field and records rapid changes due to changes in the evolution of known diseases, the apparition of new pathologies, the life expectancy of the population, progress in information theory, communication and computer sciences, development of new materials, and approaches as solutions for medical problems. The concept of personalized medicine changes the way of health care, the patient becomes the core of the decisional process, and the applied diagnostic methods and/or treatment closely fit the needs and particularities of the patient [242]. Different diagnostic or monitoring devices such as wearable health monitoring systems [243, 244], liquid biopsy or associated approaches [245, 246], wireless ultrasound transducer [247], or other point-of-care testing (POCT) methods [248, 249] are introduced and need proper analysis and validation. Furthermore, the availability of big data opens a new pathway in analyzing medical data, and artificial intelligence approaches will probably change the way of imaging diagnostic and monitoring [250, 251]. The ethical aspects must be considered [252, 253] along with valid and reliable

methods for the assessment of old and new diagnostic approaches that are required. Space for methodological improvements exists, from designing the experiments to analyzing of the experimental data for both observational and interventional approaches.

6. Concluding Remarks

Any diagnostic test falls between perfect and useless test, and no diagnostic test can tell us with certainty if a patient has or not a particular disease. No ideal diagnostic tests exist, so any test has false-positive and false-negative results.

The metric reported in the assessment of the precision (variability analysis) or accuracy of a diagnostic test must be presented as point indicators and associated 95% confidence interval, and the thresholds for interpretation are applied to the confidence intervals.

The correct evaluation of performances of two methods measuring the same outcome is done with the Bland and Altman plot (evaluate the bias of the difference between two methods) not correlation or agreement (assess the association between two measurements) analysis.

A *gold standard* test is mandatory in the evaluation of the accuracy of a test. Both sensitivity and specificity with 95% confidence intervals are reported together to allow a proper interpretation of the accuracy. Based on these values, the clinical utility index is used to support the rule-in and/or rule-out and thus respectively the usefulness of a diagnostic test as identification of the disease or in screening.

The correct interpretation of positive and negative predictive values is just made if the prevalence of the disease is known.

The sensitivity and specificity must be reported any time when Youden's index is given. Report the ROC analysis by providing AUC with associated 95% confidence interval, the threshold according to Youden's index, sensitivity, and specificity with 95% confidence intervals.

Report full descriptive and inferential statistics associated with the benefits analysis. Multivariate or multifactorial analysis could be used to test the cost-benefit of a diagnostic test, and the good practice in reporting such analysis must be strictly followed by providing the full model with the values of coefficients associated to the predictors and measures of variability, significance of both models and each coefficient, and risk metrics with associated 95% confidence intervals when appropriate (e.g., relative risk and hazard ratio).

Conflicts of Interest

The author declares that she have no conflicts of interest.

References

[1] H. Singh, "Helping health care organizations to define diagnostic errors as missed opportunities in diagnosis," *Joint Commission Journal on Quality and Patient Safety*, vol. 40, no. 3, pp. 99–101, 2014.

[2] G. D. Schiff, O. Hasan, S. Kim et al., "Diagnostic error in medicine: analysis of 583 physician-reported errors," *Archives of Internal Medicine*, vol. 169, no. 20, pp. 1881–1887, 2009.

[3] L. Zwaan and H. Singh, "The challenges in defining and measuring diagnostic error," *Diagnosis*, vol. 2, no. 2, pp. 97–103, 2015.

[4] D. L. Sackett, R. B. Haynes, G. H. Guyatt, and P. Tugwell, *Clinical Epidemiology, A Basic Science for Clinical Medicine*, Little Brown, Boston, MA, USA, 2nd edition, 1991.

[5] R. Jaeschke, G. Guyatt, and D. L. Sackett, "Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-Based Medicine Working Group," *JAMA*, vol. 271, no. 5, pp. 389–391, 1994.

[6] R. Jaeschke, G. H. Guyatt, and D. L. Sackett, "Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group," *JAMA*, vol. 271, no. 9, pp. 703–707, 1994.

[7] C. Gluud and L. L. Gluud, "Evidence based diagnostics," *BMJ*, vol. 330, pp. 724–726, 2005.

[8] EQUATOR network, Enhancing the QUAlity and Transparency of health Research, 2019, http://www.equator-network.org.

[9] C. Gamble, A. Krishan, D. Stocken et al., "Guidelines for the content of statistical analysis plans in clinical trials," *JAMA*, vol. 318, no. 23, pp. 2337–2343, 2017.

[10] J. Guilhot, M. Baccarani, R. E. Clark et al., "Definitions, methodological and statistical issues for phase 3 clinical trials in chronic myeloid leukemia: a proposal by the European LeukemiaNet," *Blood*, vol. 119, no. 25, pp. 5963–5971, 2012.

[11] J. Matcham, S. Julious, S. Pyke et al., "Proposed best practice for statisticians in the reporting and publication of pharmaceutical industry-sponsored clinical trials," *Pharmaceutical Statistics*, vol. 10, no. 1, pp. 70–73, 2011.

[12] R. Wang, S. W. Lagakos, J. H. Ware, D. J. Hunter, and J. M. Drazen, "Statistics in medicine--reporting of subgroup analyses in clinical trials," *New England Journal of Medicine*, vol. 357, no. 21, pp. 2189–2194, 2007.

[13] M. Boers, "Graphics and statistics for cardiology: designing effective tables for presentation and publication," *Heart*, vol. 104, pp. 192–200, 2018.

[14] T. A. Lang and D. G. Altman, "Basic statistical reporting for articles published in biomedical journals: the "statistical analyses and methods in the published literature" or the SAMPL guidelines," *International Journal of Nursing Studies*, vol. 52, no. 1, pp. 5–9, 2015.

[15] P. Whiting, A. Rutjes, J. Reitsma, P. Bossuyt, and J. Kleijnen, "The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews," *BMC Medical Research Methodology*, vol. 3, no. 1, 2003.

[16] P. F. Whiting, A. W. S. Rutjes, M. E. Westwood et al., "QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies," *Annals of Internal Medicine*, vol. 155, no. 8, pp. 529–536, 2011.

[17] P. M. Bossuyt, J. B. Reitsma, D. E. Bruns et al., "Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative," *Clinical Chemistry*, vol. 49, no. 1, pp. 1–6, 2003.

[18] P. M. Bossuyt, J. B. Reitsma, D. E. Bruns et al., "Standards for reporting of diagnostic accuracy," *Annals of Internal Medicine*, vol. 138, no. 1, p. W1, 2003.

[19] J. F. Cohen, D. A. Korevaar, D. G. Altman et al., "STARD 2015 guidelines for reporting diagnostic accuracy studies:

explanation and elaboration," *BMJ Open*, vol. 6, no. 11, article e012799, 2016.

[20] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. Moons, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement," *BMJ*, vol. 350, article g7594, 2015.

[21] B. J. Shea, J. M. Grimshaw, G. A. Wells et al., "Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews," *BMC Medical Research Methodology*, vol. 7, no. 1, 2007.

[22] B. J. Shea, B. C. Reeves, G. Wells et al., "AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both," *BMJ*, vol. 358, article j4008, 2017.

[23] The AGREE Collaboration. Writing Group, F. A. Cluzeau, J. S. Burgers et al., "Development and validation of an international appraisal instrument for assessing the quality of clinical practice guidelines: the AGREE project," *Quality and Safety in Health Care*, vol. 12, no. 1, pp. 18–23, 2003.

[24] M. C. Brouwers, K. Kerkvliet, and K. Spithoff, "The AGREE reporting checklist: a tool to improve reporting of clinical practice guidelines," *BMJ*, vol. 352, article i1152, 2016.

[25] Y. Chen, K. Yang, A. Marušić et al., "A reporting tool for practice guidelines in health care: the RIGHT statement," *Annals of Internal Medicine*, vol. 166, no. 2, pp. 128–132, 2017.

[26] N. L. Wilczynski, "Quality of reporting of diagnostic accuracy studies: no change since STARD statement publication--before-and-after study," *Radiology*, vol. 248, no. 3, pp. 817–823, 2008.

[27] D. A. Korevaar, W. A. van Enst, R. Spijker, P. M. Bossuyt, and L. Hooft, "Reporting quality of diagnostic accuracy studies: a systematic review and meta-analysis of investigations on adherence to STARD," *Evidence Based Medicine*, vol. 19, no. 2, pp. 47–54, 2014.

[28] L. Gallo, N. Hua, M. Mercuri, A. Silveira, and A. Worster, "Adherence to standards for reporting diagnostic accuracy in emergency medicine research," *Academic Emergency Medicine*, vol. 24, no. 8, pp. 914–919, 2017.

[29] E. N. Maclean, I. S. Stone, F. Ceelen, X. Garcia-Albeniz, W. H. Sommer, and S. E. Petersen, "Reporting standards in cardiac MRI, CT, and SPECT diagnostic accuracy studies: analysis of the impact of STARD criteria," *European Heart Journal Cardiovascular Imaging*, vol. 15, no. 6, pp. 691–700, 2014.

[30] C. Chiesa, L. Pacifico, J. F. Osborn, E. Bonci, N. Hofer, and B. Resch, "Early-onset neonatal sepsis: still room for improvement in procalcitonin diagnostic accuracy studies," *Medicine*, vol. 94, article e1230, 30 pages, 2015.

[31] Y. J. Choi, M. S. Chung, H. J. Koo, J. E. Park, H. M. Yoon, and S. H. Park, "Does the reporting quality of diagnostic test accuracy studies, as defined by STARD 2015, affect citation?," *Korean Journal of Radiology*, vol. 17, no. 5, pp. 706–714, 2016.

[32] P. J. Hong, D. A. Korevaar, T. A. McGrath et al., "Reporting of imaging diagnostic accuracy studies with focus on MRI subgroup: Adherence to STARD 2015," *Journal of Magnetic Resonance Imaging*, vol. 47, no. 2, pp. 523–544, 2018.

[33] S. Talwar, A. Sieberhofer, B. Williams, and L. Ng, "Influence of hypertension, left ventricular hypertrophy, and left ventricular systolic dysfunction on plasma *N* terminal pre-BNP," *Heart*, vol. 83, pp. 278–282, 2000.

[34] F. Corradi, C. Brusasco, A. Vezzani et al., "Hemorrhagic shock in polytrauma patients: early detection with renal doppler resistive index measurements," *Radiology*, vol. 260, no. 1, pp. 112–1128, 2011.

[35] Z. Razavi and M. Ahmadi, "Efficacy of thrice-daily versus twice-daily insulin regimens on glycohemoglobin (Hb A1c) in type 1 diabetes mellitus: a randomized controlled trial," *Oman Medical Journal*, vol. 26, no. 1, pp. 10–13, 2011.

[36] M. S. Gee, M. Atri, A. I. Bandos, R. S. Mannel, M. A. Gold, and S. I. Lee, "Identification of distant metastatic disease in uterine cervical and endometrial cancers with FDG PET/CT: analysis from the ACRIN 6671/GOG 0233 multicenter trial," *Radiology*, vol. 287, no. 1, pp. 176–184, 2018.

[37] C. M. Rerucha, R. J. Caro, and V. L. Wheeler, "Cervical cancer screening," *American Family Physician*, vol. 97, no. 7, pp. 441–448, 2018.

[38] T. Badrick, "Evidence-based laboratory medicine," *Clinical Biochemist Reviews*, vol. 34, no. 2, pp. 43–46, 2013.

[39] S. K. Rodgers, C. Chang, J. T. DeBardeleben, and M. M. Horrow, "Normal and abnormal US findings in early first-trimester pregnancy: review of the society of radiologists in ultrasound 2012 consensus panel recommendations," *RadioGraphics*, vol. 35, no. 7, pp. 2135–2148, 2015.

[40] A. Bua, P. Molicotti, S. Cannas, M. Ruggeri, P. Olmeo, and S. Zanetti, "Tuberculin skin test and QuantiFERON in children," *New Microbiologica*, vol. 36, no. 2, pp. 153–156, 2013.

[41] S. L. Liauw, S. P. Pitroda, S. E. Eggener et al., "Evaluation of the prostate bed for local recurrence after radical prostatectomy using endorectal magnetic resonance imaging," *International Journal of Radiation Oncology*, vol. 85, no. 2, pp. 378–384, 2013.

[42] American Academy of Pediatrics Committee on Fetus and Newborn and American College of Obstetricians and Gynecologists Committee on Obstetric Practice, "The Apgar score," *Pediatrics*, vol. 136, no. 4, pp. 819–822, 2015.

[43] G. Teasdale and B. Jennett, "Assessment of coma and impaired consciousness," *The Lancet*, vol. 304, no. 7872, pp. 81–84, 1974.

[44] A. Alvarado, "A practical score for the early diagnosis of acute appendicitis," *Annals of Emergency Medicine*, vol. 15, no. 5, pp. 557–564, 1986.

[45] L. M. Lenghel, C. Botar Jid, S. D. Bolboacă et al., "Comparative study of three sonoelastographic scores for differentiation between benign and malignant cervical lymph nodes," *European Journal of Radiology*, vol. 84, no. 6, pp. 1075–1082, 2015.

[46] J. E. J. Ware and C. D. Sherbourne, "The MOS 36-item short-form health survey (SF-36): I. conceptual framework and item selection," *Medical Care*, vol. 30, pp. 473–483, 1992.

[47] EuroQol Group, "EuroQol-a new facility for the measurement of health-related quality of life," *Health Policy*, vol. 16, no. 3, pp. 199–208, 1990.

[48] R. Rabin and F. de Charro, "EQ-5D: a measure of health status from the EuroQol group," *Annals of Medicine*, vol. 33, no. 5, pp. 337–343, 2001.

[49] M. B. Morgan, T. Crayford, B. Murrin, and S. C. Fraser, "Developing the vascular quality of life questionnaire: a new disease-specific quality of life measure for use in lower limb ischemia," *Journal of Vascular Surgery*, vol. 33, no. 4, pp. 679–687, 2001.

[50] J. Nordanstig, C. Wann-Hansson, J. Karlsson, M. Lundström, M. Pettersson, and M. B. Morgan, "Vascular quality of life questionnaire-6 facilitates health-related quality of life assessment in peripheral arterial disease," *Journal of Vascular Surgery*, vol. 59, no. 3, pp. 700–707, 2014.

[51] R. Launois, J. Reboul-Marty, and B. Henry, "Construction and validation of a quality of life questionnaire in chronic lower limb venous insufficiency (CIVIQ)," *Quality of Life Research*, vol. 5, no. 6, pp. 539–554, 1996.

[52] M. Korff, J. Ormel, F. J. Keefe, and S. F. Dworkin, "Grading the severity of chronic pain," *Pain*, vol. 50, pp. 133–149, 1992.

[53] R. M. Corbo and R. Scacchi, "Apolipoprotein E (apoE) allele distribution in the world. Is apoE∗4 a 'thrifty' allele?," *Annals of Human Genetics*, vol. 63, pp. 301–310, 1999.

[54] H. Boulenouar, S. Mediene Benchekor, D. N. Meroufel et al., "Impact of APOE gene polymorphisms on the lipid profile in an Algerian population," *Lipids in Health and Disease*, vol. 12, no. 1, 2013.

[55] Y. Yu, W. Wang, S. Zhai, S. Dang, and M. Sun, "IL6 gene polymorphisms and susceptibility to colorectal cancer: a meta-analysis and review," *Molecular Biology Reports*, vol. 39, no. 8, pp. 8457–8463, 2012.

[56] G. E. D. Urquhart and J. C. Gould, "Simplified technique for counting the number of bacteria in urine and other fluids," *Journal of Clinical Pathology*, vol. 18, no. 4, pp. 480–482, 1965.

[57] F. Postollec, A.-G. Mathot, M. Bernard, M.-L. Divanach, S. Pavan, and D. Sohier, "Tracking spore-forming bacteria in food: From natural biodiversity to selection by processes," *International Journal of Food Microbiology*, vol. 158, pp. 1–8, 2012.

[58] A. Özen, E. U. Unal, S. Mola et al., "Glasgow aneurysm score in predicting outcome after ruptured abdominal aortic aneurysm," *Vascular*, vol. 23, no. 2, pp. 120–123, 2015.

[59] R. A. Ciocan, C. Drugan, C. D. Gherman et al., "Evaluation of Chitotriosidase as a marker of inflammatory status in critical limb ischemia," *Annals of Clinical & Laboratory Science*, vol. 47, no. 6, pp. 713–719, 2017.

[60] C. Drugan, T. C. Drugan, N. Miron, P. Grigorescu-Sido, I. Nascu, and C. Catana, "Evaluation of neopterin as a biomarker for the monitoring of Gaucher disease patients," *Hematology*, vol. 21, no. 6, pp. 379–386, 2016.

[61] J. E. Sánchez-Rodríguez, M. Bartolomé, A. I. Cañas et al., "Anti-smoking legislation and its effects on urinary cotinine and cadmium levels," *Environmental Research*, vol. 136, pp. 227–233, 2015.

[62] N. Nahar, N. Khan, R. K. Chakraborty et al., "Color doppler sonography and resistivity index in the differential diagnosis of hepatic neoplasm," *Mymensingh Medical Journal*, vol. 23, no. 1, pp. 35–40, 2014.

[63] C. Botar Jid, S. D. Bolboacă, R. Cosgarea et al., "Doppler ultrasound and strain elastography in the assessment of cutaneous melanoma: preliminary results," *Medical Ultrasonography*, vol. 17, no. 4, pp. 509–514, 2015.

[64] M. G. Hăşmăşanu, S. D. Bolboacă, M. Matyas, and G. C. Zaharie, "Clinical and Echocardiographic Findings in Newborns of Diabetic Mothers," *Acta Clinica Croatica*, vol. 54, no. 4, pp. 458–466, 2015.

[65] BS ISO 5725-1, *Accuracy (Trueness and Precision) of Measurement Methods and Results-Part 1: General Principles and Definitions*, 1994, https://www.evs.ee/preview/iso-5725-1-1994-en.pdf.

[66] H. E. Solberg, "International Federation of Clinical Chemistry (IFCC), Scientific Committee, Clinical Section, Expert Panel on Theory of Reference Values, and International Committee for Standardization in Haematology (ICSH), Standing Committee on Reference Values. Approved Recommendation (1986) on the theory of reference values. Part 1. The concept of reference values," *Journal of Clinical Chemistry and Clinical Biochemistry*, vol. 25, no. 5, pp. 337–342, 1987.

[67] International Federation of Clinical Chemistry (IFCC). Scientific Committee, Clinical Section. Expert Panel on Theory of Reference Values (EPTRV). IFCC Document (1982) stage 2, draft 2, 1983-10-07 with a proposal for an IFCC recommendation. The theory of reference values. Part 2. Selection of individuals for the production of reference values," *Clinica Chimica Acta*, vol. 139, no. 2, pp. 205F–213F, 1984.

[68] H. E. Solberg and C. PetitClerc, "International Federation of Clinical Chemistry (IFCC), Scientific Committee, Clinical Section, Expert Panel on Theory of Reference Values. Approved recommendation (1988) on the theory of reference values. Part 3. Preparation of individuals and collection of specimens for the production of reference values," *Journal of Clinical Chemistry and Clinical Biochemistry*, vol. 26, no. 9, pp. 593–598, 1988.

[69] H. E. Solberg and D. Stamm, "IFCC recommendation: The theory of reference values. Part 4. Control of analytical variation in the production, transfer and application of reference values," *Journal of Automat Chemistry*, vol. 13, no. 5, pp. 231–234, 1991.

[70] H. E. Solberg, "Approved recommendation (1987) on the theory of reference values. Part 5. Statistical treatment of collected reference values. Determination of reference limits," *Journal of Clinical Chemistry and Clinical Biochemistry*, vol. 25, pp. 645–656, 1987.

[71] R. Dybkær and H. E. Solberg, "Approved recommendation (1987) on the theory of reference values. Part 6. Presentation of observed values related to reference values," *Journal of Clinical Chemistry and Clinical Biochemistry*, vol. 25, pp. 657–662, 1987.

[72] K. Thygesen, J. S. Alpert, A. S. Jaffe et al., "Third universal definition of myocardial infarction," *Journal of the American College of Cardiology*, vol. 60, no. 16, pp. 1581–1598, 2012.

[73] R. Weitgasser, B. Gappmayer, and M. Pichler, "Newer portable glucose meters--analytical improvement compared with previous generation devices?," *Clinical Chemistry*, vol. 45, no. 10, pp. 1821–1825, 1999.

[74] D. B. Sacks, M. Arnold, G. L. Bakris et al., "Guidelines and recommendations for laboratory analysis in the diagnosis and management of diabetes mellitus," *Diabetes Care*, vol. 34, no. 6, pp. e61–e99, 2011.

[75] A. Clerico, A. Ripoli, S. Masotti et al., "Evaluation of 99th percentile and reference change values of a high-sensitivity cTnI method: A multicenter study," *Clininica Chimica Acta*, vol. 493, pp. 156–161, 2019.

[76] M. Le, D. Flores, D. May, E. Gourley, and A. K. Nangia, "Current practices of measuring and reference range reporting of free and total testosterone in the United States," *Journal of Urology*, vol. 195, no. 5, pp. 1556–1561, 2016.

[77] A. B. Alnor and P. J. Vinholt, "Paediatric reference intervals are heterogeneous and differ considerably in the classification of healthy paediatric blood samples," *European Journal of Pediatrics*, 2019, In press.

[78] C. McGee, A. Hoehn, C. Hoenshell, S. McIlrath, H. Sterling, and H. Swan, "Age- and gender-stratified adult myometric reference values of isometric intrinsic hand strength," *Journal of Hand Therapy*, no. 18, pii: S0894-1130, article 30352, 2019.

[79] X. T. Zhang, K. M. Qi, J. Cao, K. L. Xu, and H. Cheng, "[Survey and Establishment of the Hematological Parameter Reference Intervals for Adult in Xuzhou Area of China],"

Zhongguo Shi Yan Xue Ye Xue Za Zhi, vol. 27, no. 2, pp. 549–556, 2019.

[80] K. Adeli, V. Higgins, M. Nieuwesteeg et al., "Biochemical marker reference values across pediatric, adult, and geriatric ages: establishment of robust pediatric and adult reference intervals on the basis of the Canadian Health Measures Survey," *Clinical Chemistry*, vol. 61, pp. 1049–1062, 2015.

[81] O. Addai-Mensah, D. Gyamfi, R. V. Duneeh et al., "Determination of haematological reference ranges in healthy adults in three regions in Ghana," *BioMed Research International*, vol. 2019, Article ID 7467512, 2019.

[82] A. Li, S. Yang, J. Zhang, and R. Qiao, "Establishment of reference intervals for complete blood count parameters during normal pregnancy in Beijing," *Journal of Clinical Laboratory Analysis*, vol. 31, no. 6, article e22150, 2017.

[83] G. Siest, J. Henny, R. Gräsbeck et al., "The theory of reference values: an unfinished symphony," *Clinical Chemistry and Laboratory Medicine*, vol. 51, no. 1, pp. 47–64, 2013.

[84] Y. Ozarda, K. Sikaris, T. Streichert, J. Macri, and IFCC Committee on Reference intervals and Decision Limits (C-RIDL), "Distinguishing reference intervals and clinical decision limits-A review by the IFCC committee on reference intervals and decision limits," *Critical Reviews in Clinical Laboratory Sciences*, vol. 55, no. 6, pp. 420–431, 2018.

[85] S. D. Colan, "The why and how of *Z* scores," *Journal of the American Society of Echocardiography*, vol. 26, no. 1, pp. 38–40, 2013.

[86] A. Field, *An Adventure in Statistics: The Reality Enigma*, pp. 189–214, Sage, London, UK, 2016.

[87] A. Hazra and N. Gogtay, "Biostatistics series module 7: the statistics of diagnostic tests," *Indian Journal of Dermatology*, vol. 62, no. 1, pp. 18–24, 2017.

[88] H. Chubb and J. M. Simpson, "The use of *Z*-scores in paediatric cardiology," *Annals of Pediatric Cardiology*, vol. 5, no. 2, pp. 179–184, 2012.

[89] A. E. Curtis, T. A. Smith, B. A. Ziganshin, and J. A. Elefteriades, "The mystery of the *Z*-score," *Aorta*, vol. 4, no. 4, pp. 124–130, 2016.

[90] Y. K. Mao, B. W. Zhao, L. Zhou, B. Wang, R. Chen, and S.S. Wang, "*Z*-score reference ranges for pulsed-wave doppler indices of the cardiac outflow tracts in normal fetuses," *International Journal of Cardiovascular Imaging*, vol. 35, no. 5, pp. 811–825, 2019.

[91] C. L. Gregson, S. A. Hardcastle, C. Cooper, and J. H. Tobias, "Friend or foe: high bone mineral density on routine bone density scanning, a review of causes and management," *Rheumatology*, vol. 52, no. 6, pp. 968–985, 2013.

[92] T. B. Newman, W. S. Browner, S. R. Cummings, and S. B. Hulley, "Designing studies of medical tests," in *Designing Clinical Research*, S. B. Hulley, S. R. Cummings, W. S. Browner, D. G. Grady, and T. B. Newman, Eds., pp. 171–191, Lippincott Williams & Wilkins, Phyladelphia, PA, USA, 4th edition, 2013.

[93] N. Taheri, G. Roshandel, M. Mojerloo et al., "Comparison of serum levels of hepcidin and pro-hepcidin in hemodialysis patients and healthy subjects," *Saudi Journal of Kidney Diseases and Transplantation*, vol. 26, no. 1, pp. 34–38, 2015.

[94] J. Zhang, Y. Zhao, and Y. Chen, "Reference intervals for plasma pro-gastrin releasing peptide (ProGRP) levels in healthy adults of Chinese Han ethnicity," *International Journal of Biological Markers*, vol. 29, no. 4, pp. e436–e439, 2014.

[95] S. H. Saravelos and T. C. Li, "Intra- and inter-observer variability of uterine measurements with three-dimensional ultrasound and implications for clinical practice," *Reproductive Biomedicine Online*, vol. 31, no. 4, pp. 557–564, 2015.

[96] C. Z. Azara, E. J. Manrique, N. L. Alves de Souza, A. R. Rodrigues, S. B. Tavares, and R. G. Amaral, "External quality control of cervical cytopathology: interlaboratory variability," *Acta Cytologica*, vol. 57, no. 6, pp. 585–590, 2013.

[97] J. Lei, P. Yang, L. Zhang, Y. Wang, and K. Yang, "Diagnostic accuracy of digital breast tomosynthesis versus digital mammography for benign and malignant lesions in breasts: a meta-analysis," *European Radiology*, vol. 24, no. 3, pp. 595–602, 2014.

[98] D. L. Miglioretti, E. Johnson, A. Williams et al., "The use of computed tomography in pediatrics and the associated radiation exposure and estimated cancer risk," *JAMA Pediatrics*, vol. 167, no. 8, pp. 700–707, 2013.

[99] T. Dinh, U. Ladabaum, P. Alperin, C. Caldwell, R. Smith, and T. R. Levin, "Health benefits and cost-effectiveness of a hybrid screening strategy for colorectal cancer," *Clinical Gastroenterology and Hepatology*, vol. 11, no. 9, pp. 1158–1166, 2013.

[100] J. F. Tiernan, S. Gilhooley, M. E. Jones et al., "Does an interferon-gamma release assay change practice in possible latent tuberculosis?," *QJM*, vol. 106, no. 2, pp. 139–146, 2013.

[101] T. B. Newman, W. S. Browner, S. R. Cummings, and S. B. Hulley, "Chapter 12. Designing studies of medical tests," in *Designing Clinical Research*, S. B. Hulley, S. R. Cummings, W. S. Browner, D. G. Grady, and T. B. Newman, Eds., pp. 170–191, Lippincott Williams & Wilkins, Philadelphia, PA, USA, 4th edition, 2007.

[102] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3–4, pp. 591–611, 1965.

[103] A. N. Kolmogorov, "Sulla determinazione empirica di una legge di distribuzione," *Giornale dell'Istituto Italiano degli Attuari*, vol. 4, pp. 83–91, 1933.

[104] N. Smirnov, "Table for estimating the goodness of fit of empirical distributions," *Annals of Mathematical Statistics*, vol. 19, pp. 279–281, 1948.

[105] W. N. Arifin, A. Sarimah, B. Norsaadah et al., "Reporting statistical results in medical journals," *Malaysian Journal of Medical Science*, vol. 23, no. 5, pp. 1–7, 2016.

[106] J. L. Peacock, S. M. Kerry, and R. R. Balise, "Chapter 5. Introduction to presenting statistical analysis," in *Presenting Medical Statistics from Proposal to Publication*, pp. 48–51, Oxford University Press, UK, 2nd edition, 2017.

[107] A. Koch and H. Singer, "Normal values of B type natriuretic peptide in infants, children, and adolescents," *Heart*, vol. 89, no. 8, pp. 875–878, 2003.

[108] CLSI, Defining, *Establishing, and Verifying Reference Intervals in the Clinical Laboratory; Approved Guideline—Third Edition, CLSI document EP28-A3c*, Clinical and Laboratory Standards Institute, Wayne, PA, USA, 2008, https://clsi.org/media/1421/ep28a3c_sample.pdf.

[109] T. Murase, H. Kitamura, T. Kochi et al., "Distributions and ranges of values of blood and urinary biomarker of inflammation and oxidative stress in the workers engaged in office machine manufactures: evaluation of reference values," *Clinical Chemistry and Laboratory Medicine*, vol. 51, no. 2, pp. 421–428, 2013.

[110] Calculating Inter- and Intra-Assay Coefficients of Variability, 2018, https://www.salimetrics.com/calculating-inter-and-intra-assay-coefficients-of-variability/.

[111] T. Chard, *An Introduction to Radioimmunoassay and Related Techniques*, Elsevier, Amsterdam, Netherlands, 1995.
[112] D. Wild, *The Immunoassay Handbook: Theory and Applications of Ligand Binding, ELISA and Related Techniques*, Elsevier, Amsterdam, Netherlands, 4th edition, 2013.
[113] K. Pearson, "Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity, and Panmixia," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 187, pp. 253–318, 1896.
[114] R. G. D. Steel and J. H. Torrie, *Principles and Procedures of Statistics*, McGraw-Hill, New York, NY, USA, 2nd edition, 1980.
[115] P. Sangnawakij and S. Niwitpong, "Confidence intervals for coefficients of variation in two-parameter exponential distributions," *Communications in Statistics-Simulation and Computation*, vol. 46, no. 8, pp. 6618–6630, 2016.
[116] B. Everitt, *The Cambridge Dictionary of Statistics*, Cambridge University Press, Cambridge, UK, 1998, ISBN 0521593468.
[117] J. H. Zar, *Biostatistical Analysis*, p. 32, Prentice Hall, Inc., Englewood Cliffs, NJ, USA, 2nd edition, 1984
[118] D. Zwillinger and S. Kokoska, *Standard Probability and Statistical Tables and Formula*, p. 17, Chapman & Hall, Boca Raton, FL, USA, 2000.
[119] D. G. Bonett, "Confidence interval for a coefficient of quartile variation," *Computational Statistics & Data Analysis*, vol. 50, no. 11, pp. 2953–2957, 2006.
[120] Đ. T. N. Quyên, "Developing university governance indicators and their weighting system using a modified Delphi method," *Procedia-Social and Behavioral Sciences*, vol. 141, pp. 828–833, 2014.
[121] J. Forkman, "Estimator and tests for common coefficients of variation in normal distributions," *Communications in Statistics-Theory and Methods*, vol. 38, pp. 233–251, 2009.
[122] C. J. Feltz and G. E. Miller, "An asymptotic test for the equality of coefficients of variation from k populations," *Statistics in Medicine*, vol. 15, no. 6, pp. 647–658, 1996.
[123] K. Krishnamoorthy and M. Lee, "Improved tests for the equality of normal coefficients of variation," *Computational Statistics*, vol. 29, no. 1-2, pp. 215–232, 2014.
[124] B. Marwick and K. Krishnamoorthy, "cvequality: Tests for the Equality of Coefficients of Variation from Multiple Groups," *R Software Package Version 0.1.3*, 2018, https://github.com/benmarwick/cvequality.
[125] A. L. Schafer, E. Vittinghoff, R. Ramachandran, N. Mahmoudi, and D. C. Bauer, "Laboratory reproducibility of biochemical markers of bone turnover in clinical practice," *Osteoporosis International*, vol. 21, no. 3, pp. 439–445, 2010.
[126] J. L. Calvi, F. R. Chen, V. B. Benson et al., "Measurement of cortisol in saliva: a comparison of measurement error within and between international academic-research laboratories," *BMC Research Notes*, vol. 10, no. 1, 2017.
[127] G. F. Reed, F. Lynn, and B. D. Meade, "Use of coefficient of variation in assessing variability of quantitative assays," *Clinical and Diagnostic Laboratory Immunology*, vol. 9, no. 6, pp. 1235–1239, 2002.
[128] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.
[129] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.
[130] A. J. Conger, "Integration and generalization of kappas for multiple raters," *Psychological Bulletin*, vol. 88, no. 2, pp. 322–328, 1980.
[131] J. Cohen, "Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit," *Psychological Bulletin*, vol. 70, no. 4, pp. 213–220, 1968.
[132] K. Pearson, "Notes on regression and inheritance in the case of two parents," *Proceedings of the Royal Society of London*, vol. 58, no. 347-352, pp. 240–242, 1895.
[133] J. J. Bartko, "The intraclass correlation coefficient as a measure of reliability," *Psychological Reports*, vol. 19, no. 1, pp. 3–11, 1966.
[134] L. I.-K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, no. 1, pp. 255–268, 1989.
[135] L. I.-K. Lin, "A note on the concordance correlation coefficient," *Biometrics*, vol. 56, pp. 324-325, 2000.
[136] D. G. Altman and J. M. Bland, "Measurement in medicine: the analysis of method comparison studies," *Statistician*, vol. 32, pp. 307–317, 1983.
[137] J. M. Bland and D. G. Altman, "Measuring agreement in method comparison studies," *Statistical Methods in Medical Research*, vol. 8, pp. 135–160, 1999.
[138] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia Medica*, vol. 22, no. 3, pp. 276–282, 2012.
[139] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of Chiropractic Medicine*, vol. 15, no. 2, pp. 155–163, 2016.
[140] K. O. McGraw and S. P. Wong, "Forming inferences about some intraclass correlation coefficients," *Psychological Methods*, vol. 1, pp. 30–46, 1996.
[141] G. B. McBride, "A proposal for strength-of-agreement criteria for Lin's concordance correlation coefficient," NIWA Report HAM2005-062, 2018, https://www.medcalc.org/download/pdf/McBride2005.pdf.
[142] W. P. Martins and C. O. Nastri, "Interpreting reproducibility results for ultrasound measurements," *Ultrasound in Obstetrics & Gynecology*, vol. 43, no. 4, pp. 479-480, 2014.
[143] S. M. Gregoire, U. J. Chaudhary, M. M. Brown et al., "The Microbleed Anatomical Rating Scale (MARS): reliability of a tool to map brain microbleeds," *Neurology*, vol. 73, no. 21, pp. 1759–1566, 2009.
[144] T. H. Lee, J. S. Lee, S. J. Hong et al., "High-resolution manometry: reliability of automated analysis of upper esophageal sphincter relaxation parameters," *Turkish Journal of Gastroenterolpgy*, vol. 25, no. 5, pp. 473–480, 2014.
[145] V. Abdollah, E. C. Parent, and M. C. Battié, "Is the location of the signal intensity weighted centroid a reliable measurement of fluid displacement within the disc?," *Biomedizinische Technik. Biomedical Engineering*, vol. 63, no. 4, pp. 453–460, 2018.
[146] J. W. Cuchna, M. C. Hoch, and J. M. Hoch, "The interrater and intrarater reliability of the functional movement screen: a systematic review with meta-analysis," *Physical Therapy in Sport*, vol. 19, pp. 57–65, 2016.
[147] N. Parenti, M. L. Reggiani, P. Iannone, D. Percudani, and D. Dowding, "A systematic review on the validity and reliability of an emergency department triage scale, the Manchester Triage System," *International Journal of Nursing Studies*, vol. 51, no. 7, pp. 1062–1069, 2014.
[148] T. Lange, O. Matthijs, N. B. Jain, J. Schmitt, J. Lützner, and C. Kopkow, "Reliability of specific physical examination tests for the diagnosis of shoulder pathologies: a systematic review

and meta-analysis," *British Journal of Sports Medicine*, vol. 51, no. 6, pp. 511–518, 2017.

[149] J. M. Bland and D. G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *The Lancet*, vol. 327, no. 8476, pp. P307–310, 1986.

[150] D. Giavarina, "Understanding Bland Altman analysis," *Biochemia Medica*, vol. 25, no. 2, pp. 141–151, 2015.

[151] J. S. Krouwer, "Why Bland-Altman plots should use X, not $(Y+X)/2$ when X is a reference method," *Statistics in Medicine*, vol. 27, pp. 778–780, 2008.

[152] G. Montalescot, U. Sechtem, S. Achenbach et al., "2013 ESC guidelines on the management of stable coronary artery disease: the Task Force on the management of stable coronary artery disease of the European Society of Cardiology," *European Heart Journal*, vol. 34, no. 38, pp. 2949–3003, 2013.

[153] T. Pincus and T. Sokka, "Laboratory tests to assess patients with rheumatoid arthritis: advantages and limitations," *Rheumatic Disease Clinics of North America*, vol. 35, no. 4, pp. 731–734, 2009.

[154] A. Aboraya, C. France, J. Young, K. Curci, and J. LePage, "The validity of psychiatric diagnosis revisited," *Psychiatry*, vol. 2, no. 9, pp. 48–55, 2005.

[155] M. A. den Bakker, "[Is histopathology still the gold standard?]," *Ned Tijdschr Geneeskd*, vol. 160, article D981, 2017.

[156] M. M. Leeflang, P. M. Bossuyt, and L. Irwig, "Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis," *Journal of Clinical Epidemiology*, vol. 62, no. 1, pp. 5–12, 2009.

[157] M. M. Leeflang, J. J. Deeks, Y. Takwoingi, and P. Macaskill, "Cochrane diagnostic test accuracy reviews," *Systematic Reviews*, vol. 2, no. 1, 2013.

[158] S. D. Bolboacă, L. Jäntschi, A. F. Sestraş, R. E. Sestraş, and D. C. Pamfil, "Pearson-Fisher chi-square statistic revisited," *Information*, vol. 2, no. 3, pp. 528–545, 2011.

[159] W. J. Youden, "Index for rating diagnostic tests," *Cancer*, vol. 3, pp. 32–35, 1950.

[160] A. S. Glas, J. G. Lijmer, M. H. Prins, G. J. Bonsel, and P. M. Bossuyt, "The diagnostic odds ratio: a single indicator of test performance," *Journal of Clinical Epidemiology*, vol. 6, no. 11, pp. 1129–1135, 2003.

[161] M. J. Galloway and M. M. Reid, "Is the practice of haematology evidence based? III. Evidence based diagnostic testing," *Journal of Clinical Pathology*, vol. 51, pp. 489–491, 1998.

[162] F. Habibzadeh and M. Yadollahie, "Number needed to misdiagnose: a measure of diagnostic test effectiveness," *Epidemiology*, vol. 24, no. 1, p. 170, 2013.

[163] A. J. Mitchell, "The clinical significance of subjective memory complaints in the diagnosis of mild cognitive impairment and dementia: a meta-analysis," *International Journal of Geriatric Psychiatry*, vol. 23, no. 11, pp. 1191–1202, 2008.

[164] A. J. Mitchell, "Sensitivity × PPV is a recognized test called the clinical utility index (CUI+)," *European Journal of Epidemiology*, vol. 26, no. 3, pp. 251-252, 2011.

[165] T. J. Fagan, "Nomogram for Bayes theorem," *New England Journal of Medicine*, vol. 293, no. 5, p. 257, 1975.

[166] C. G. B. Caraguel and R. Vanderstichel, "The two-step Fagan's nomogram: ad hoc interpretation of a diagnostic test result without calculation," *BMJ Evidence-Based Medicine*, vol. 18, no. 4, pp. 125–128, 2013.

[167] J. Marasco, R. Doerfler, and L. Roschier, "Doc, what are my chances," *UMAP Journal*, vol. 32, pp. 279–298, 2011.

[168] M. H. Zweig and G. Campbell, "Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine," *Clinical Chemistry*, vol. 39, no. 4, pp. 561–577, 1993.

[169] C. E. Metz, "ROC methodology in radiologic imaging," *Investigative Radiology*, vol. 21, no. 9, pp. 720–733, 1986.

[170] T. A. Lasko, J. G. Bhagwat, K. H. Zou, and L. Ohno-Machado, "The use of receiver operating characteristic curves in biomedical informatics," *Journal of Biomedical Informatics*, vol. 38, no. 5, pp. 404–415, 2005.

[171] H. Ma, A. I. Bandos, H. E. Rockette, and D. Gur, "On use of partial area under the ROC curve for evaluation of diagnostic performance," *Statistics in Medicine*, vol. 32, no. 20, pp. 3449–3458, 2013.

[172] F. Habibzadeh, P. Habibzadeh, and M. Yadollahie, "On determining the most appropriate test cut-off value: the case of tests with continuous results," *Biochemia Medica*, vol. 26, no. 3, pp. 297–307, 2016.

[173] K. Hajian-Tilaki, "The choice of methods in determining the optimal cut-off value for quantitative diagnostic test evaluation," *Statistical Methods in Medical Research*, vol. 27, no. 8, pp. 2374–2383, 2018.

[174] A. R. Chiorean, M. B. Szep, D. S. Feier, M. Duma, M. A. Chiorean, and Ş. Strilciuc, "Impact of strain elastography on BI-RADS classification in small invasive lobular carcinoma," *Medical Ultrasonography*, vol. 18, no. 2, pp. 148–153, 2018.

[175] J. V. Carter, J. Pan, S. N. Rai, and S. Galandiuk, "ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves," *Surgery*, vol. 159, no. 6, pp. 1638–1645, 2016.

[176] K. O. Hajian-Tilaki, A. R. Gholizadehpasha, S. Bozorgzadeh, and E. Hajian-Tilaki, "Body mass index and waist circumference are predictor biomarkers of breast cancer risk in Iranian women," *Medical Oncology*, vol. 28, no. 4, pp. 1296–1301, 2011.

[177] H. Javidi, "Psychometric properties of GHQ-12 in Goa, India," *Asian Journal of Psychiatry*, vol. 30, p. 141, 2017.

[178] How Good is that Test? II [online] © copyright 1994–2007, Bandolier–"Evidence-based thinking about health care," 2018, http://www.bandolier.org.uk/band27/b27-2.html.

[179] F. Habibzadeh, "How to report the results of public health research," *Journal of Public Health and Emergency*, vol. 1, p. 90, 2017.

[180] D. W. Dippel, A. de Kinkelder, S. L. Bakker, F. van Kooten, H. van Overhagen, and P. J. Koudstaal, "The diagnostic value of colour duplex ultrasound for symptomatic carotid stenosis in clinical practice," *Neuroradiology*, vol. 41, no. 1, pp. 1–8, 1999.

[181] A.-M. Demers, S. Verver, A. Boulle et al., "High yield of culture-based diagnosis in a TB-endemic setting," *BMC Infectious Diseases*, vol. 12, no. 1, 2012.

[182] F. H. Schröder, J. Hugosson, M. J. Roobol et al., "The European randomized study of screening for prostate cancer–prostate cancer mortality at 13 years of follow-up," *Lancet*, vol. 384, no. 9959, pp. 2027–2035, 2014.

[183] J. Hugosson, R. A. Godtman, S. V. Carlsson et al., "Eighteen-year follow-up of the Göteborg randomized population-based prostate cancer screening trial: effect of sociodemographic variables on participation, prostate cancer incidence and mortality," *Scandinavian Journal of Urology*, vol. 52, no. 1, pp. 27–37, 2018.

[184] F. Attili, C. Fabbri, I. Yasuda et al., "Low diagnostic yield of transduodenal endoscopic ultrasound-guided fine needle

biopsy using the 19-gauge flex needle: a large multicenter prospective study," *Endoscopic Ultrasound*, vol. 6, no. 6, pp. 402–408, 2017.

[185] R. Gulati, A. B. Mariotto, S. Chen, J. L. Gore, and R. Etzioni, "Long-term projections of the number needed to screen and additional number needed to treat in prostate cancer screening," *Journal of Clinical Epidemiology*, vol. 64, no. 12, pp. 1412–1417, 2011.

[186] J. Hugosson, S. Carlsson, G. Aus et al., "Mortality results from the Göteborg randomised population-based prostate-cancer screening trial," *Lancet Oncology*, vol. 11, no. 8, pp. 725–732, 2010.

[187] M. S. Lee, J. Y. Cho, S. Y. Kim et al., "Diagnostic value of integrated PET/MRI for detection and localization of prostate cancer: Comparative study of multiparametric MRI and PET/CT," *Journal of Magnetic Resonance Imaging*, vol. 45, no. 2, pp. 597–609, 2017.

[188] P. Mozafari, R. M. Azari, Y. Shokoohi, and M. Sayadi, "Feasibility of biological effective monitoring of chrome electroplaters to chromium through analysis of serum malondialdehyde," *International Journal of Occupational and Environmental Medicine*, vol. 7, no. 4, pp. 199–206, 2016.

[189] C. Lertudomphonwanit, R. Mourya, L. Fei et al., "Large-scale proteomics identifies MMP-7 as a sentinel of epithelial injury and of biliary atresia," *Science Translational Medicine*, vol. 9, no. 417, article eaan8462, 2017.

[190] G. Porta, F. G. Numis, V. Rosato et al., "Lactate determination in pleural and abdominal effusions: a quick diagnostic marker of exudate-a pilot study," *Internal and Emergency Medicine*, vol. 13, no. 6, pp. 901–906, 2018.

[191] J. T. Bram, K. D. Baldwin, and T. J. Blumberg, "Gram stain is not clinically relevant in treatment of pediatric septic arthritis," *Journal of Pediatric Orthopaedics*, vol. 38, no. 9, pp. e536–e540, 2018.

[192] A. L. Shetty, T. Brown, T. Booth et al., "Systemic inflammatory response syndrome-based severe sepsis screening algorithms in emergency department patients with suspected sepsis," *Emergency Medicine Australasia*, vol. 28, pp. 287–294, 2016.

[193] A. J. Mitchell, J. B. McGlinchey, D. Young, I. Chelminski, and M. Zimmerman, "Accuracy of specific symptoms in the diagnosis of major depressive disorder in psychiatric outpatients: data from the MIDAS project," *Phychological Medicine*, vol. 39, pp. 1107–1116, 2009.

[194] M. M. Johansson, A. S. Kvitting, E. Wressle, and J. Marcusson, "Clinical utility of cognistat in multiprofessional team evaluations of patients with cognitive impairment in Swedish primary care," *International Journal of Family Medicine*, vol. 2014, Article ID 649253, 10 pages, 2014.

[195] A. J. Mitchell, M. Yadegarfar, J. Gill, and B. Stubbs, "Case finding and screening clinical utility of the patient health questionnaire (PHQ-9 and PHQ-2) for depression in primary care: a diagnostic meta-analysis of 40 studies," *British Journal of Psychiatry Open*, vol. 2, pp. 127–138, 2016.

[196] S.-M. Fereshtehnejad, , M. Shafieesabet, A. Rahmani, et al., "A novel 6-item screening questionnaire for parkinsonism: validation and comparison between different instruments," *Neuroepidemiology*, vol. 43, pp. 178–193, 2014.

[197] F. Bartoli, C. Crocamo, E. Biagi et al., "Clinical utility of a single-item test for DSM-5 alcohol use disorder among outpatients with anxiety and depressive disorders," *Drug and Alcohol Dependence*, vol. 165, pp. 283–287, 2016.

[198] K. Hoti, M. Atee, and J. D. Hughes, "Clinimetric properties of the electronic Pain Assessment Tool (ePAT) for aged-care residents with moderate to severe dementia," *Journal of Pain Research*, vol. 2018, no. 11, pp. 1037–1044, 2018.

[199] A. J. Mitchell, D. Shukla, H. A. Ajumal, B. Stubbs, and T. A. Tahir, "The mini-mental state examination as a diagnostic and screening test for delirium: systematic review and meta-analysis," *General Hospital Psychiatry*, vol. 36, no. 6, pp. 627–633, 2014.

[200] M. Gothlin, M. Eckerstrom, S. Rolstad, P. Kettunen, and A. Wallin, "Better prognostic accuracy in younger mild cognitive impairment patients with more years of education," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 10, pp. 402–412, 2018.

[201] K. V. Fernandes, A. Jhobta, M. Sarkar, S. Thakur, and R. G. Sood, "Comparative evaluation of 64- slice multidetector CT virtual bronchoscopy with fiberoptic bronchoscopy in the evaluation of tracheobronchial neoplasms," *International Journal of Medical and Health Research*, vol. 3, no. 8, pp. 40–47, 2017.

[202] E. Cugy and I. Sibon, "Stroke-associated pneumonia risk score: validity in a French stroke unit," *Journal of Stroke and Cerebrovascular Diseases*, vol. 26, no. 1, pp. 225–229, 2017.

[203] S. D. Pawar, J. D. Naik, P. Prabhu, G. M. Jatti, S. B. Jadhav, and B. K. Radhe, "Comparative evaluation of Indian diabetes risk score and Finnish diabetes risk score for predicting risk of diabetes mellitus type II: A teaching hospital-based survey in Maharashtra," *Journal of Family Medicine and Primary Care*, vol. 6, pp. 120–125, 2017.

[204] S. Gur-Ozmen, A. Leibetseder, H. R. Cock, N. Agrawal, and T. J. von Oertzen, "Screening of anxiety and quality of life in people with epilepsy," *Seizure*, vol. 45, pp. 107–113, 2017.

[205] J. Scott, S. Marwaha, A. Ratheesh et al., "Bipolar at-risk criteria: an examination of which clinical features have optimal utility for identifying youth at risk of early transition from depression to bipolar disorders," *Schizophrenia Bulletin*, vol. 43, no. 4, pp. 737–744, 2017.

[206] R. Albadareen, G. Gronseth, M. Goeden, M. Sharrock, C. Lechtenberg, and Y. Wang, "Paraneoplastic autoantibody panels: sensitivity and specificity, a retrospective cohort," *International Journal of Neuroscience*, vol. 127, no. 6, pp. 531–538, 2017.

[207] J. D. B. Diestro, P. M. D. Pasco, L. V. Lee, and XDP Study Group of the Philippine Children's Medical Center, "Validation of a screening questionnaire for X-linked dystonia parkinsonism: The first phase of the population-based prevalence study of X-linked dystonia parkinsonism in Panay," *Neurology and Clinical Neuroscience*, vol. 5, pp. 79–85, 2017.

[208] C. Dawes, *Oxford–Center for Evidence-Based Medicine*, 2019, https://www.cebm.net/2014/06/catmaker-ebm-calculators/.

[209] R. Herbert, *Confidence Interval Calculator*, 2019, https://www.pedro.org.au/english/downloads/confidence-interval-calculator/.

[210] C. Caraguel, R. Wohlers-Reichel, and R. Vanderstichel, "DocNomo: the app to add evidence to your diagnosis," in *Proceedings of the Australian Veterinary Association Annual Conference*, Adelaide, Australia, May 2016.

[211] J. Allen, S. Graziadio, and M. Power, *A Shiny Tool to Explore Prevalence, Sensitivity, and Specificity on Tp, Fp, Fn, and Tn*, NIHR Diagnostic Evidence Co-operative Newcastle, Newcastle upon Tyne, UK, 2017, https://micncltools.shinyapps.io/TestAccuracy/.

[212] M. Power, S. Graziadio, and J. Allen, *A ShinyApp Tool To Explore Dependence of Rule-In And Rule-Out Decisions on*

Prevalence, Sensitivity, Specificity, and Confidence Intervals, NIHR Diagnostic Evidence Co-operative Newcastle, Newcastle upon Tyne, UK, 2017, https://micncltools.shinyapps.io/ClinicalAccuracyAndUtility/.

[213] T. R. Fanshawe, M. Power, S. Graziadio, J. M. Ordóñez-Mena, J. Simpson, and J. Allen, "Interactive visualisation for interpreting diagnostic test accuracy study results," *BMJ Evidence-Based Medicine*, vol. 23, no. 1, pp. 13–16, 2018.

[214] A. I. Mushlin, H. S. Ruchlin, and M. A. Callahan, "Cost effectiveness of diagnostic tests," *The Lancet*, vol. 358, no. 9290, pp. 1353–1355, 2001.

[215] M. Drummond, "Economic evaluation of health interventions," *BMJ*, vol. 337, article a1204, 2008.

[216] G. Pennello, N. Pantoja-Galicia, and S. Evans, "Comparing diagnostic tests on benefit-risk," *Journal of Biopharmaceutical Statistics*, vol. 26, no. 6, pp. 1083–1097, 2016.

[217] A. J. Vickers, B. Van Calster, and E. W. Steyerberg, "Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests," *BMJ*, article i6, 2016.

[218] S. Bolboacă and L. Jäntschi, "Pearson versus Spearman, Kendall's tau correlation analysis on structure-activity relationships of biologic active compounds," *Leonardo Journal of Sciences*, vol. 9, pp. 179–200, 2006.

[219] L. Jäntschi, D. Balint, and S. D. Bolboacă, "Multiple linear regressions by maximizing the likelihood under assumption of generalized Gauss-Laplace distribution of the error," *Computational and Mathematical Methods in Medicine*, vol. 2016, Article ID 8578156, 8 pages, 2016.

[220] J. L. Peacock, S. M. Kerry, and R. R. Basile, *Presenting Medical Statistics from Proposal to Publication*, Oxford University Press, Oxford, UK, 2nd edition, 2017.

[221] S. D. Bolboacă and L. Jäntschi, "Sensitivity, specificity, and accuracy of predictive models on phenols toxicity," *Journal of Computational Science*, vol. 5, no. 3, pp. 345–350, 2014.

[222] S. D. Bolboacă and L. Jäntschi, "The effect of leverage and/or influential on structure-activity relationships," *Combinatorial Chemistry & High Throughput Screening*, vol. 16, no. 4, pp. 288–297, 2013.

[223] S. D. Bolboacă and L. Jäntschi, "Quantitative structure-activity relationships: linear regression modelling and validation strategies by example," *International Journal on Mathematical Methods and Models in Biosciences*, vol. 2, no. 1, article 1309089, 2013.

[224] S. D. Bolboacă and L. Jäntschi, "Distribution fitting 3. analysis under normality assumptions," *Bulletin of University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca Horticulture*, vol. 66, no. 2, pp. 698–705, 2009.

[225] S. D. Bolboacă and L. Jäntschi, "Modelling the property of compounds from structure: statistical methods for models validation," *Environmental Chemistry Letters*, vol. 6, pp. 175–181, 2008.

[226] L. Schlessinger and D. M. Eddy, "Archimedes: a new model for simulating health care systems—the mathematical formulation," *Journal of Biomedical Informatics*, vol. 35, no. 1, pp. 37–50, 2002.

[227] J. Horsman, W. Furlong, D. Feeny, and G. Torrance, "The Health Utilities Index (HUI®): concepts, measurement properties and applications," *Health and Quality of Life Outcomes*, vol. 1, no. 1, p. 54, 2003.

[228] M. Westwood, T. van Asselt, B. Ramaekers et al., "High-sensitivity troponin assays for the early rule-out or diagnosis of acute myocardial infarction in people with acute chest pain: a systematic review and cost-effectiveness analysis," *Health Technology Assessment*, vol. 19, no. 44, pp. 1–234, 2015.

[229] S. Daya, "Study design for the evaluation of diagnostic tests," *Seminars in Reproductive Endocrinology*, vol. 14, no. 2, pp. 101–109, 1996.

[230] S. P. Glasser, "Research methodology for studies of diagnostic tests," in *Essentials of Clinical Research*, S. P. Glasser, Ed., pp. 245–257, Springer, Dordrecht, Netherlands, 2008.

[231] J. G. Lijmer, B. W. Mol, S. Heisterkamp et al., "Empirical evidence of design-related bias in studies of diagnostic tests," *JAMA*, vol. 282, no. 11, pp. 1061–1066, 1999.

[232] Centre for Evidence Based Medicine, 2019, https://www.cebm.net/.

[233] T. McGinn, R. Jervis, J. Wisnivesky, S. Keitz, P. C. Wyer, and Evidence-based Medicine Teaching Tips Working Group, "Tips for teachers of evidence-based medicine: clinical prediction rules (CPRs) and estimating pretest probability," *Journal of General Internal Medicine*, vol. 23, no. 8, pp. 1261–1268, 2008.

[234] W. S. Richardson, M. C. Wilson, S. A. Keitz, P. C. Wyer, and EBM Teaching Scripts Working Group, "Tips for teachers of evidence-based medicine: making sense of diagnostic test results using likelihood ratios," *Journal of General Internal Medicine*, vol. 23, no. 1, pp. 87–92, 2008.

[235] S. Carley, S. Dosman, S. Jones, and M. Harrison, "Simple nomograms to calculate sample size in diagnostic studies," *Emergency Medicine Journal*, vol. 22, no. 3, pp. 180-181, 2005.

[236] E. W. de Bekker-Grob, B. Donkers, M. F. Jonker, and E. A. Stolk, "Sample size requirements for discrete-choice experiments in healthcare: a practical guide," *Patient*, vol. 8, no. 5, pp. 373–384, 2015.

[237] M. A. Bujang and T. H. Adnan, "Requirements for minimum sample size for sensitivity and specificity analysis," *Journal of Clinical & Diagnostic Research*, vol. 10, no. 10, pp. YE01–YE06, 2016.

[238] K. Hajian-Tilaki, "Sample size estimation in diagnostic test studies of biomedical informatics," *Journal of Biomedical Informatics*, vol. 48, pp. 193–204, 2014.

[239] G. Realdi, L. Previato, and N. Vitturi, "Selection of diagnostic tests for clinical decision making and translation to a problem oriented medical record," *Clinica Chimica Acta*, vol. 393, no. 1, pp. 37–43, 2008.

[240] C. S. Kosack, A. L. Page, and P. R. Klatser, "A guide to aid the selection of diagnostic tests," *Bulletin of the World Health Organization*, vol. 95, no. 9, pp. 639–645, 2017.

[241] A. M. Buehler, B. de Oliveira Ascef, H. A. de Oliveira Júnior, C. P. Ferri, and J. G. Fernandes, "Rational use of diagnostic tests for clinical decision making," *Revista da Associação Médica Brasileira*, vol. 65, no. 3, pp. 452–459, 2019.

[242] M. Di Sanzo, L. Cipolloni, M. Borro et al., "Clinical applications of personalized medicine: a new paradigm and challenge," *Current Pharmaceutical Biotechnology*, vol. 18, no. 3, pp. 194–203, 2017.

[243] Y. Liu, H. Wang, W. Zhao, M. Zhang, H. Qin, and Y. Xie, "Flexible, stretchable sensors for wearable health monitoring: sensing mechanisms, materials, fabrication strategies and features," *Sensors*, vol. 18, no. 2, p. 645, 2018.

[244] J.-F. Masson, "Surface plasmon resonance clinical biosensors for medical diagnostics," *ACS Sensors*, vol. 2, no. 1, pp. 16–30, 2017.

[245] G. Méhes, "Liquid biopsy for predictive mutational profiling of solid cancer: the pathologist's perspective," *Journal of Biotechnology*, vol. 297, pp. 66–70, 2019.

[246] X. Ma, S. He, B. Qiu, F. Luo, L. Guo, and Z. Lin, "Noble metal nanoparticle-based multicolor immunoassays: an approach toward visual quantification of the analytes with the naked eye," *ACS Sensors*, vol. 4, no. 4, pp. 782–791, 2019.

[247] R. Haridy, "Pocket-sized, affordably-priced ultrasound connects to an iPhone," *News Atlas*, 2017, https://newatlas.com/butterfly-iq-smartphone-ultrasound/51962/.

[248] J. Katoba, D. Kuupiel, and T. P. Mashamba-Thompson, "Toward improving accessibility of point-of-care diagnostic services for maternal and child health in low-and middle-income countries," *Point of Care*, vol. 18, no. 1, pp. 17–25, 2019.

[249] Y. Gong, Y. Zheng, B. Jin et al., "A portable and universal upconversion nanoparticle-based lateral flow assay platform for point-of-care testing," *Talanta*, vol. 201, no. 15, pp. 126–133, 2019.

[250] J. H. Thrall, X. Li, Q. Li et al., "Artificial intelligence and machine learning in radiology: opportunities, challenges, pitfalls, and criteria for success," *Journal of the American College of Radiology*, vol. 15, no. 3, pp. 504–508, 2018.

[251] B. Gallix and J. Chong, "Artificial intelligence in radiology: who's afraid of the big bad wolf?," *European Radiology*, vol. 29, no. 4, pp. 1637–1639, 2019.

[252] D. W. Dowdy, C. R. Gounder, E. L. Corbett, L. G. Ngwira, R. E. Chaisson, and M. W. Merritt, "The ethics of testing a test: randomized trials of the health impact of diagnostic tests for infectious diseases," *Clinical Infectious Diseases*, vol. 55, no. 11, pp. 1522–1526, 2012.

[253] A. E. Bulboacă, S. D. Bolboacă, and A. C. Bulboacă, "Ethical considerations in providing an upper limb exoskeleton device for stroke patients," *Medical Hypotheses*, vol. 101, pp. 61–64, 2017.