The identification of evidence supporting inadequate science in contemporary reference ranges and a need for advancement. A step toward unambiguous clinical diagnostic tools. An Overview review

A Thesis submitted in partial fulfillment of the requirements of the Doctor of Osteopathy degree program

Submitted To: NATIONAL UNIVERSITY OF MEDICAL SCIENCES

Submitted By: Matthew Paul Gibbons

Student I.D Number: S140914

Submitted: June 9/2016

Table of Contents

List of figures Page

Introduction

Modern medicine heavily relies on laboratory diagnostic testing with nearly 80% of decisions based on clinical assays (Forsman 1996, Utah Governor Huntsman 2009, Hanson and Plumhoff 2012, Morehouse 2013, Hallworth 2014, Wolcott et al 2014, Muennig, Akhmetov et al 2015) and reliance on them continues to increase (Mindemark et al 2011, Hauser & Shirts 2014). A retrospective analysis from 1993 through 2013 discovered that in vitro diagnostic testing had an annual growth rate in the United States of 5.3%, increasing from US$30 billion in 1998 to an estimated US$67 billion in 2013 (Rohr et al 2016, Beastall 2013). European *in vitro* diagnostics test expenditure range from €3.6 in Romania to €43.5 in Switzerland per capita per annum (European IVD 2014). As of 2014 over $1 trillion USD in healthcare is wasted with a large portion of this coming from the improper use of laboratory testing (Berwick and Hackbarth 2012, Bulger et al 2013). Between 7 and 10 billion lab tests are performed annually (Futrell 2015), with statistics showing anywhere between 20-50% of the time testing being inapt to the patient's condition (Verbrugghe et al 2014, Morehouse 2103, van Walraven and Raymond 2003, Miyakis et al 2006, Fushimi et al 2006, Rollins 2012, Carter 2014, Naugler 2014, Liu et al 2012). Out of over 500 million patients per year who have lab tests performed as part of their assessment, a potential 23 million are alarm for significant concern due to inaccurate correlation between disease and lab results (Hickner 2014). Anywhere amid 40,000-80,000 U.S hospital deaths occur from misdiagnosis per year (Newman-Toker 2009). Even when the proper tests are carried out under proper conditions much dysregulation is not properly identified. This is largely due to the incomplete science of reference range guidelines in which subjects are compared to the "healthy population." The aperture in evidence-based practice results in reduced diagnostic accuracy affecting patient prognosis, further adding to a superfluous economic burden (Lippi & Mattiuzzi 2013). We can economically improve medical expenses (Waters et al 2011). Value-based healthcare is defined as "Maximizing outcomes over cost by moving away from fee for service models to ones that reward providers on the basis of outcomes" (St John et al 2015, Porter & Teisberg 2015). Limited evidence-based for the effectiveness of diagnostic services is well known as well as little evidence of cost effectiveness (St John and Price 2013). The incidence of misleading laboratory test results epitomizes the need for studying and improving laboratory utilization (Hogg et al 2005, Walter et al 2013, Hauser and Shirts 2014). With such dependence on these values, it is imperative that the diagnostic accuracy continues to improve. The objective of this paper is to emphasize the need to update the exactitude of reference ranges (RR).

Methods

A literature search in the PUBMED and SCI-HUB databases was conducted from December 2015 to February 2016 was performed using the free text words: "Reference Ranges" OR "Laboratory diagnostics" AND “Healthy” were used in combination. These resources were used because they most likely to contain relevant information. Reading related topics from a personal reference collection identified the keywords.

The search was limited to English and human studies. This reference lists from the articles obtained was then cross-referenced with the articles already obtained. When the search results began to produce no new references the search was terminated. Relevant articles from the 1950’s on (beginning of laboratory diagnostics) were used.

The inclusion criteria were:

- Systematic reviews, prospective controlled trials, retrospective and prospective cohort, case-control, cross-sectional, case series / case study, expert opinion or narrative reviews
- Related to the science of clinical laboratory diagnostics and reference ranges

The exclusion criteria were:

- Studies without the above

Results

The initial search using the words “Reference Range” OR “Laboratory Diagnostic Accuracy” and "Healthy" identified 2,730 articles. No further new information occurred after 217 papers were reviewed therefore the review was stopped. These 217 papers were then sorted based on the inclusion criteria leaving 129 articles that met the inclusion criteria. 88 papers did not meet the inclusion criteria.

History of Reference Ranges

Since the 1960’s assay results are based on reference ranges (Grasbeck and Saris 1969, Siest et al 2013, Hauser and Shirts 2014). The initial twenty years saw the development of several biological variability themes. Then from 1980 onward international recommendations from multiple reputable sources such as IFCC-LM (International Federation of Clinical Chemistry and Laboratory Medicine)(Thienpont et al 2013), scientific societies [French (SFBC), Spanish (SEQC)(Siest et al 2013), Canadian Laboratory Initiative on Pediatric Reference Intervals (CALIPER) and many more began to publish reference interval guidelines (Adeli et al 2015). These are available and organized in textbooks and of several congresses, workshops, and round

tables all over the world. At the beginning of the millennia several concepts for universal RR were proposed by several groups (Siest et al 2013).

Need for change

It is a high priority for modern medicine to improve upon the accuracy of the existing reference ranges largely because they do not necessarily confirm that you have a disease if you fall outside these "normal" values (Boyd 2010, Katayev et al 2010, Walter et al 2013). Evidence presented in this review emphasizes how factors such as regional population differences (Ichihara et al 2004, Ichihara et al 2008), individual factors from person to person (even with similar health profiles), and clinical application of evidence-based laboratory medicine influence how RR validity is lacking fruition. The basis for the effectiveness of any analytic test must be that it is both dependable and precise in its clinical conclusion (Zidan et al 2013). It is imperative and overdue that international collaboration ensues reviewing the validity and reliability to define a new set of standards. A review by Christopher P Price gives a clear definition of evidence-based laboratory medicine (EBLM), which states 'the conscientious, judicious and explicit use of best evidence in the use of laboratory medicine investigations for assisting in making decisions about the care of individual patients' (Price 2012). This was derived from the definition of Evidence-Based Medicine given by Sackett *et al 1996.*

Establishing reference ranges

Critical appraisal of the literature of biological markers will perpetually reveal the infinite dynamic between the biomarker and pathology. Unfortunately this knowledge does not reveal how the test and ranges may be used for individual patients (Rector et al 2012). Sackett and Haynes proposed that a hierarchy of evidence was necessary to establish the evidence for use of a diagnostic test. Keeping in mind that evidence alone is not enough to ensure optimization; a prevailing set of guidelines must be put into action (Sackett and Haynes 2002).

Specifications for developing and classifying RR involve samples from at least 120 specimens from a healthy population and then identifying the standard deviation of 2.5% from the outermost 5% to use in defining limits for two-sided or one-side intervals (Katayev et al 2010, Horowitz et al 2010, Boyd 2010). The chemistry of the body changes immensely throughout the different stages of life significantly reflected in clinical testing. RR are established by a non parametric means for each category including age, diet, gender, circadian rhythm, race, posture, medications, physical activity, socioeconomic status, medical history, and fasting status (Huma et al 2013, Blankenstein 2015). The population whose laboratory results will be compared to this reference range should demographically match (Wener 2011). The quality of the reference ranges can play as significant a role in result interpretation, as the quality of the result itself (Phillips 2009).

According to the National Association of Testing Authorities, Australia (NATA) Field Application Document for ISO 15189 section 5.5.5 and reads as follows: *The sources of biological reference intervals and/or medical decision points must be documented and should include references to the information used in deciding the intervals, any statistical processes used, literature studies considered and the personnel involved in deciding the intervals. Where possible and relevant, customers of the laboratory with appropriate expertise should also be involved in the determination of reference intervals. Consideration should be given to adopting intervals/decision points consistent with those in other laboratories, where possible and appropriate.*

While individual laboratories use their own RR some countries use established international RR while others use domestic intervals developed by organizations over years using their population focusing on geographical location. Using vitamin D levels as an example, RR based on foreign population can be misleading. Over 60% of the population is vitamin D deficient according to international reference ranges, but the cause may be a change in the population range, ecological factors, and socio-economic (Khan et al 2013, Huma et al 2013). The absence of universal values creates challenges, especially for research facilities to create their own particular reference values. It has also lead to the unsatisfactory quality of new innovations, such as HPLC, GCMS, and PCR tests. Clinicians have developed practice rules and presented their own particular guidelines for following, decision limits, likelihood ratios, and Reference Change Value (RCV) (Deeks 2001,Siest et al 2013). This, however, complicates communication among laboratorians and clinicians in substituting reference values and decision limits in lab reports (Siest et al 2013). The acronym "SCIENCE": standardization and harmonization; clinical effectiveness; innovation; evidence-based practice; novel applications; cost-effectiveness; and education of others is a brilliant framework put in order to improve use of medical resources and patient outcomes (Beastall 2013).

It is essential that RR are defined, they act, as a tool so there is a set understanding of "Normal." These are defined as the set of values in which 95% of the normal healthy population falls (Walton 2001, Katayev et al 2010). Accurate comprehension about biomarker fluctuation concentrations throughout life and between sexes is vital to clinical interpretation of laboratory test results in different disease states (Panteghini 2004, Ricos et al 2009, Biswis et al 2015). In more recent years epidemiological outcome analysis has added decision limits to improve accuracy. For acquisition of data a carefully selected and defined reference population is vital for the intended test; it is easily possible that the specific group of individuals selected may not be representative of that population. Volunteers may also be influenced to participate in the studies for their own health concerns using the free resource as an opportunity resulting in biased population values. The goal is to determine the expected range of inter-individual variation; it is clear that if the contributions from the first two are relatively large, they will obscure the part of the total variation that is due to actual differences among individuals. Establishing a reference range is simply taking steps to reduce the magnitude of this obscuring effect, unfortunately the values are often founded on defunct methods (Boyd 2010).

Recommended elements of a process for establishing a reference interval (Jones and Barker 2008):

- Define the analyte (measure and) for which the reference interval is being established, the clinical utility, biological variation and major variations in form.
- Define the method used, the accuracy base, and analytical specificity.
- Define important pre-analytical considerations together with any actions in response to the interference.
- Define the principle behind the reference interval (i.e. central 95% etc.)
- Describe the data source(s), including: number of subjects, nature of subjects, exclusions, pre-analytical factors, statistical measures, outliers excluded and analytical method.
- Define considerations of partitioning based on age, sex etc.
- Define the number of significant figures, i.e. the degree of rounding.
- Define the clinical relevance of the reference limits.
- Consider the use of common reference intervals.
- Decision and implementation.

The original method used to study specimens was used for many years; Single-analyte assays or low-to-mid-plex procedures involved analyzing a single or low number of biomarkers. Innovative progression now allows for multiplex assays, best explained as the use of devices that simultaneously measures multiple analytes in a single run/cycle of the same specimen of the assay (commonly used for cytokines) (Wener 2011). In a recent issue of *Arthritis Research & Therapy*, Chandra and colleagues investigated the use of multiple multiplex assays (a 'megaplex'?) in the evaluation and categorization of patients with rheumatoid arthritis.

Figure 1. Criteria for evaluating multiplex assays (Chandra et al 2011)

1. Analytical performance parameters (for example, precision, analytical sensitivity, and linearity) of assays should be available for each analyte.
2. The clinical performance (clinical sensitivity and specificity) of each analyte within the multiplex should be comparable to that of assays for individual analytes.
3. The time required to produce all results of multiplex assay should be less than the sum of time required to produce results of individual assays.
4. The combination of multiplex assays should be appropriate for answering clinical questions; that is, the combinations of analytes measured within multiplex assay should make clinical sense.

Figure 2. Criteria for evaluating multivariate index assays

1. Normal control and disease control groups used to generate the index should be matched demographically (age, gender, race, and geography) with the target group.
2. Pre-analytical variables (specimen type, specimen handling, and specimen storage) should be equivalent in control and diseased groups.
3. There should be a high ratio of subjects (patients) to measured analytes used to generate the index.

4. The accuracy (clinical sensitivity and specificity) of the index test should be tested and reported on the basis of populations of subjects (the 'test set' of diseased patients and controls) independently of the subjects (the 'training set') used to generate the index formulae or calculations.
5. The clinical accuracy (clinical sensitivity and specificity) of the index test should be compared with the accuracy of the most accurate of the individual analytes within the index or with the best available single diagnostic laboratory test or both.

It is extremely difficult for laboratories to follow these recommendations to an appropriately high standard because of the numerous aspects the RR entail (Jones and Barker 2008). This merits clinical consultation with experienced clinicians to aid in establishment (Wener 2011). Proposal of a new construct for establishing RR using highly valid sources instead of unreliable patient population is an undeniable necessity.

Science of Reference Ranges

Diagnostic accuracy is primarily represented by two measures, sensitivity and specificity (Lalkhen et al 2008). Measures such as predictive values, odds-ratios, likelihood ratios, Youden's index, diagnostic odds ratio (DOR), summary receiver operating characteristic (ROC) curves, (Šimundić 2008) Inter-observer variation, and Intra-observer variation are often effective enhancements (Deeks 2001, Kanchanaraksa 2008). The data attained from a diagnostic test will frequently fall on a spectrum (i.e. blood pressure, hormone concentration), requiring a decision on whether a certain test value indicates that the condition is present (positive test) or not (negative test). This 'point' is termed the decision or positivity threshold, such as blood pressure cut-off value for hypertension (135/80).

Sensitivity refers to the probability of a person with the condition of interest having a positive result (also known as the true positive proportion [TPP]), while **specificity** is the probability of a person without the condition of interest having a negative result (also known as the true negative proportion [TNP]) (Altman 1994, Eusebi 2013, The Joanna Briggs Institute 2015).

Figure 3.

Calculation for specificity $\frac{\text{True negatives}}{\text{(True negatives+False positives)}}$

Calculation for sensitivity $\frac{\text{True positives}}{\text{(True positives+False negatives)}}$

Sensitivity and specificity measure the accuracy of a diagnostic test but do not provide the

probability of the diagnostic value of the result of the test. The idea of predictive values is to provide the proportion of patients who are correctly diagnosed (Altman 1994, Eusebi 2013, The Joanna Briggs Institute 2015).

Figure 4.

Positive $\frac{PPV=TP}{(TP+FP)}$ Negative $\frac{NPV=TN}{(TN+FN)}$

Disease prevalence correlates to predictive values, meaning the higher the prevalence the higher the positive predictive value (Mariska et al 2013).

Likelihood ratios were introduced into medicine in the late 70's. They assess probability that the result acquired would be expected in a person with the condition when compared to the probability that the same result would be in a person without. Specificity and sensitivity are used to determine whether a test result usefully changes the likelihood that a condition exists. Two forms of the likelihood ratio exist, LR+ for positive test (how many times more likely a patient with the condition will present a positive result) and LR- (how many times more likely a negative test result will be present for a patient without the condition) (Boyd 2010, Eusebi 2013).

Figure 5.

$$LR- = \frac{(1\text{-sensitivity})}{(\text{specificity})} = \frac{FN}{(TP+FN)} \div \frac{TN}{(FP+TN)}$$

A priori is dependent on the A posteriori and interpretation requires a calculator to convert between probabilities and odds of the disorder (McGee 2002,The Joanna Briggs Institute 2015).

Receiver Operating Characteristic (ROC) curve analysis is a statistical method with practical implications for evaluating the performance of diagnostic tests that classify individuals into categories of those with and those without a condition ((Metz 1978, Zou et al 2007, Eusebi 2013 The Joanna Briggs Institute 2015).

Youden's index

Youden's index is one of the eldest measures for diagnostic accuracy (Youden 1950, Okeh and Okoro 2012). It is internationally accepted as the prime measure of a tests performance. Its discriminative power of a diagnostic procedure and also is key for comparison of a test with others. Calculation of Youden's index starts by deducting 1 from the sum of test's sensitivity and specificity expressed not as percentage but as a part of a whole number: (sensitivity + specificity) – 1. Tests with poor diagnostic accuracy have a Youden's index value of 0. Then for a perfect

test YI equals 1. YI is not sensitive for differences in the sensitivity and specificity of the test a core disadvantage. Predominantly, a test with sensitivity 0,9 and specificity 0,4 has the same Youden's index (0,3) as a test with sensitivity 0,6 and specificity 0,7. It is extremely well defined that those tests are not of comparable diagnostic accuracy. If one is to assess the discriminative power of a test solely based on YI it could be incorrectly concluded by a clinician that these two tests are equally effective. A strong point of YI is that it is not affected by the disease prevalence. Unfortunately is affected by the spectrum of the disease, sensitivity specificity, likelihood ratios and DOR (López-Ratón 2016).

Diagnostic odds ratio (DOR)

Diagnostic odds ratio is another globally-implemented measure for diagnostic accuracy. It enables general estimation of discriminative power of diagnostic procedures and also for the comparison of diagnostic accuracies between two or more diagnostic tests. DOR of a test is defined as the ratio of the odds of positivity in subjects with disease relative to the odds in subjects without disease (Altman 1994, Eusebi 2013). DOR relies considerably on the sensitivity and specificity of a test. Assays with a high specificity and sensitivity combined with a low rate of false positives and false negatives have high a DOR. Just as substantial a detail, with the same sensitivity of the test, DOR increases with the increase of the test specificity; assay sensitivity > 90% and specificity of 99% has a DOR greater than 500. Similarly to sensitivity and specificity DOR depends on criteria used to define disease and its spectrum of pathological conditions of the examined group (disease severity, phase, stage, comorbidity etc.). Markedly DOR does not depend on disease prevalence (Eusebi 2013).

Figure 6. DOR analyzed according to the formula: $DOR = (TP/FN)/(FP/TN)$

Figure 7.

Index Test Outcome	Reference positive	Reference negative	Total
Index test positive (T+)	True positives (TP)	False positives (FP)	Test positives (TP+FP)
Index test negative (T-)	False negatives (FN)	True Negatives (TN)	Test negatives (FN+TN)
Total	Reference positives (TP+FN)	Reference negatives (FP+TN)	N (TP+FP+FN+TN)

Sensitivity and specificity co-vary with the decision threshold used to identify the disorder. (Lalkhen and McCluskey 2008, The Joanna Briggs Institute 2015)

Inter-observer variation is a variation in the result of a test due to multiple observers examining the result (inter = between).

Intra-observer variation is a variation in the result of a test due to the same observer examining the result at different times (intra = within)

The difference is due to the extent to which observer(s) agree or disagree when interpreting the same test result (Kanchanaraksa 2008).

Following the evidence

Evidence-based medicine is intent on implying practical application of scientific information retrieved from the research to appropriate fields. Just as importantly, it evaluates the quality of evidence relevant to the risks and benefits of individuals' characteristics or treatments by categorization and ranking according to the strength of the lack from various biases. Meta-analyses of randomized, double-blind, controlled clinical trials are considered the strongest evidence for therapeutic interventions, followed by case reports and expert opinion as the lower value (Elstein 2004, Panagiotakos 2008).

The U.S. Preventive Services Task Force ranks scientific evidence in the following order: (a)Evidence obtained from more than one randomized controlled trials (Level I); (b)Evidence obtained from controlled trials without randomization (Level II-1); or Evidence obtained from prospective or case-control epidemiologic studies (Level II-2); or Evidence obtained from multiple time series with or without the intervention (Level II-3); (c)Opinions of respected authorities, based on clinical experience, descriptive studies, or reports of expert committees (Level III).

The UK National Health Service uses a comparable system with classes labeled A, B, C, and D. Any time a selection must be made amongst numerous alternative options, a decision is being made, and the role of the researcher is to aid in this process. Significantly when decisions are complex and require cautious consideration and systematic review of the available information, the researcher's role becomes paramount (Panagiotakos 2008, Beastall 2013).

Determining diagnostic test accuracy

Many variables can affect the clinical relevance of a laboratory test. Diagnostic test accuracy studies compare a diagnostic test of interest (the 'index test') to an existing diagnostic test (the 'reference test'), known to be the best test currently available for accurately identifying the presence or absence of the condition of interest. Results are compared with one another in order to evaluate the accuracy of the index test. (The-systematic-review-of-studies-of-diagnostic-test-accuracy)

Two main study types for diagnostic test accuracy.

1. Diagnostic case- control design, also sometimes called the 'two gate design'. The first population is known to have the condition (i.e. a health care centre), and the second without the condition. This design gives an indication of the maximum accuracy of the test. The results, however, the results will generally give an inflated indication of the test's accuracy in practice.
2. Cross-sectional, all patients suspected of having the condition of interest experience both the index test and the reference test. Positive results for the condition on the reference test can be considered to be the cases, while those who test negative are the controls. This

design is understood to imitate actual practice better and is more likely to provide a valid estimate of diagnostic accuracy.

Problems with Reference Ranges

Lippi et al 2009 state that laboratory diagnostics are a vital part of clinical decision making, but is no safer than other areas of healthcare and that despite implementations to improve patient safety, we still lack concrete evidence that healthcare safety and quality of have grasped their pinnacle.

Accuracy may be adequate in delivering evidence of improvement or equivalence in patient outcomes when a well-defined target condition linked to effective downstream management consequences, such as effective treatment (Lijmer et al 2009). The accuracy model, however, is somewhat problematic, particularly whenever a new test leads to a classification in disease for which there is no clinical reference range or when the new test is thought to be better than the current. There are policies in place to deal with cases in which the RR result is missing or when information can be used to build a substitute or proxy, but when there is no accepted RR, other tactics have to be used. Sometimes a reference standard is not available, nor is it obvious how the target condition should be defined (Lord et al 2006, Bossuyt 2008, Lord et al 2009, Lippi et al 2009). There is also confusion because The Clinical Laboratory Improvement Amendments of 1988 (CLIA) does not specifically use the term "validation" but refers to "establishment of performance specifications.' These were established to bolster federal oversight of clinical laboratories to ensure the accuracy and reliability of patient test results (Code of Federal Regulations 2009, Burd 2010)

Systematic reviews of diagnostic test accuracy summarize test performance based on all available evidence, evaluate the quality of published studies, and account for discrepancy in findings between studies (Ochodo 2012). Approximations of test accuracy regularly vary between studies because of differences in how test positivity is defined, study design, patient characteristics and positioning of the test in the diagnostic pathway. Additionally, diagnostic test accuracy studies have distinctive design characteristics, which require different criteria for critical appraisal compared to other sources of quantitative evidence. They also report a pair of related summary statistics ('sensitivity and specificity', as discussed below) rather than a single statistic such as an odds ratio. Complicating the systematic reviews of these studies is the requirement of different statistical methods for meta-analytical pooling, and different approaches for narrative synthesis. While test evaluations in literature has increased there is still inconsistency with methodology. Multiple surveys have indicated only a small amount of studies follow crucial criteria for standards (Whiting 2004, Reitsma et al 2005, Reitsma et al 2012). An updated version of original QUADAS (Quality Assessment of Diagnostic Accuracy Studies) for systematic reviews has been designed through subjective experience allowing for a more concise tool known as the QUADAS-2. The four main aspects of this format are: patient selection, index test, reference standard, and flow and timing. Evaluation of each is based on risk of bias, applicability. Initially published in 2003, the QUADAS tool has been widely used in more than 200 review abstracts in the Database of Abstracts of Reviews of Effects (DARE), and cited over

500 times. For its contributions to improvement of patient outcomes the QUADAS-2 is endorsed by the Agency for Healthcare Research and Quality, Cochrane Collaboration (Leeflang and Deeks et al 2013), and the U.K. National Institute for Health and Clinical Excellence (Whiting 2011, Willis 2011).

Diagnostic accuracy studies compare results from one or more tests to the excepted reference, which is a crucial step for evaluating new and existing data (Guyatt 1986, Sackett 2002). Several aspects threaten the internal and external validity of these studies (Sheps and Schechter 1984, Begg 1987, Jaeschke et l 1994 1, Jaeschke et al 1994 2, Reid et al 1995, Mower 1999, Bossuyt et al A and B 2003*,):* study design, patient selection, testing procedure, and examination of data. An example of faulty validity from meta-analysis correlates design flaws with embellished diagnostic accuracy. This improper understanding allows for unjust acceptance of assays leading to poor treatment protocols. Assessors of these studies should take note of the possibility for bias and absence of clinical usefulness. In order to reduce the errors that may arise from this the Standards for Reporting of Diagnostic Accuracy (STARD) initiative was created. This initiative allowed for objective reporting of studies of diagnostic accuracy. The checklist was designed to bring awareness to clinicians of factors that diminish efficacy and determining of test reliability (Bossuyt 1 et al 2003, Korevaar et al 2015). Hunink, Kresin and other authors doubt the importance of test accuracy in test evaluations, arguing that the results are frequently too late to impact management and policy decisions due to the speed at which technology advances (Hunink and Krestin 2002).

How the ranges of specific biomarkers are discovered is one aspect, but the more important matter is what they actually represent. They are still not at a truly acceptable level of precision to justify accepting them as definitive. According to Boyd, “The statistical definition of the reference interval may not allow certain clinical uses. As a specific example, reference intervals are statistically derived with respect to only the healthy population; they cannot be used to rule in or rule out specific conditions such as male infertility” (Boyd 2010). Clinical trials implemented to demonstrate a beneficial effect on patients should have clinical measure as the primary outcome and to focus on the ultimate goals of healthcare: restore or maintain health, survival, activity level, function, and reduce disability. Some trials and studies do not have this in mind as the priority, sometimes focusing on other outcome measures such as resources, length of stay, satisfaction, or results and not consequences. The mainstay for testing leading to improved patient outcome is through changes in clinical decision-making and guidelines developed by the test results. These guidelines include selecting, starting, stopping, or modifying treatment; ordering more tests; or attentive monitoring (Bossuyt and McCaffery 2009).

Foundational controversy

One argument in the efficacy of RR is the emphasis on the range of the average population rather than people with an overall ideal level of health know as optimal (O'Donnell 1987). Optimal health range or therapeutic target (not to be confused with biological target) is defined a reference range or limit that is based on concentrations or levels that are associated with optimal health or minimal risk of related complications and diseases, rather than the standard range based on normal distribution in the population. The issues with this idea are first; in sample populations

used to find these intervals the individuals chosen may not be at ideal levels of health and give a false view of norm. Secondly, a fundamental flaw with optimal health range is the lack of standardized method for estimating the ranges; there are many variations and interpretations of "optimal" depending on the source. Because of this, educated estimates are used to decide what is considered optimal for each individual (National Committee for Clinical Laboratory Standards 1995, LaboratorySM 2001, Yadav et al 2015) 'When you examine the 2 test results from difference populations you will promptly realize that what is normal for one group is not essentially normal for another group' (Huma et al 2013). The members of a group may show a varying degree of sings and symptoms from no occult to overt related the systems that are being tested. Just because an individual falls in the normal range does not mean they do not have a dysregulation, or visa versa and qualify as healthy if they are in the healthy population range. "Health is a relative condition lacking a universal definition. Defining what is considered healthy becomes the initial problem in any study...." (Wayne 2008) Your individual reference range — "will often be a better barometer of disease risk than a score on a reference range established by testing others," Kaufman said. At times results differ so considerably within the population that the laboratory may reference a smaller proportion of the population (Chicago Tribune 2011). For instance, the RR normally quoted for serum insulin may only include results within one standard deviation above and one standard deviation below the mean value. This includes 68% of the reference population. In this case, 16% of normal people will have 'abnormal' high insulin and 16% will have 'abnormal' low insulin according to the quoted reference range. Serum insulin is therefore not a useful test for assessing 'insulin resistance' (Phillips 2009).

Normal people with abnormal results

There are several major reasons why a healthy individual can produced test results showing an abnormal comparison to the accepted intervals. Laboratory technology allows for multiple biochemical analysis to be done by one machine and produce up to 20 results. These results are not independent and possibly come from abnormally disturbed reference populations. This means only approximately 36% of normal people will have all 20 results in the RR leaving 64% with a minimum of one abnormal result. The more extreme an abnormal result and the more similar tests are abnormal, shows this abnormality is of clinical relevance. Statistical values verifying this can be understood by considering that 99% reference range (approx. ± 2.6 standard deviations) and the 99.9% reference range (approx. ± 3.3 standard deviations), 82% and 98% of people will have all 20 tests within the RR (0.99^{20} and 0.999^{20} respectively). This is valuable for understanding an isolated abnormal result. To illustrate quoting from *(Aust Prescr 2009;32:43–6)*. "For example, the reference range of alkaline phosphatase is 30–110 U/L. This covers two standard deviations below the mean and two above the mean. One standard deviation is therefore 20 U/L [(110–30) ÷ 4]. A result of 150 U/L is two standard deviations above the upper limit of the reference range and therefore four standard deviations above the mean. " This is not likely to occur in a normal individual but this result could be normal if the RR it was based on was inaccurate. Determining result abnormality requires consideration of similar tests. Alkaline phosphatase is one of several 'LFTs'-liver function tests (others include bilirubin, gamma glutamyl transferase, alanine aminotransferase, aspartate aminotransferase and lactate dehydrogenase). Irregularities in comparable tests would insinuate that the abnormal alkaline

phosphatase could be the result of liver disease, whilst elevated alkaline phosphatase in isolation may imply another problem, for example bone pathology (Phillips 2009).

Other examples of RR inaccuracy include evidence of subclinical autoimmune thyroid disease prevalence. It is estimated at up to 40% of woman whom show lymphocytic infiltration of the thyroid and 10-15% with autoantibodies. It is so common that seemingly the diseased population could easily contaminate healthy individuals (Dayan 1996). The US National Academy of Clinical Biochemistry (NACB) recommends the use of a revised normal range for thyroid disease. If only persons negative for antibodies against thyroid peroxidase and with no personal history of thyroid pathology are tested, 95% of TSH values lie within 0.48–3.60 (Bjoro et al 2000). Several studies have discovered an increase in thyroid peroxidase antibody positivity with TSH concentrations outlier the narrow range 0·2–1·9 mU/L, offering evidence that TSH in the upper reference range often accompanies abnormal pathology in the thyroid (Michalopoulou et al 1998, Bjoro et al 2000, Hak et al 2000, Dayan 2002, Baloch 2013).

Individual variation in reference ranges

Aside from discrepancies during specimen collection, differing laboratory methods and other potential factors the greatest possibility for variability is within the patient themselves (Panteghini 2004, Phillips 2009). Age, sex, hormone, diurnal and seasonal cycles, behavioural, nutrition and several other variables heavily influence the results of an assay. For continuous monitoring the time factor has the greatest impact with longer time frames allowing for greater extremes. No matter what measures clinicians and laboratorians take to reduce the distorting effects of these influences it can only be controlled to a certain degree (Phillips 2009). With such a reliance on the accuracy of the RR, how are we to determine if all of these factors are not giving a completely obscured indication of an individuals health? This leads to a very important aspect about the changes to a patient during treatment: are the changes monitored in values coming from the treatment or from intra-individual variability?

In statistics, '**regression to the mean**' is the phenomenon that if a variable is extreme on its first measurement, it will tend to be closer to the average on its second measurement, if it is extreme on its second measurement it will tend to have been closer to the average on its first (Irwig 1991, Phillips 2009).

With this phenomenon, however, the homogeneous patients will probably fall within the 95% of RR in one of the first rounds of testing; upon further testing, however, the extremes could vary each time, falling in and out of this standard deviation. Normally healthy patients will fall within the RR and continue to do so upon continuous testing, however with this phenomenon the homogeneous patients will probably fall within the 95% RR one the first round of testing but upon further testing the extremes could varying each time falling in and out of this standard deviation. The preliminary outcomes at the extremes are the result of extreme random variability in one direction or the other. A similar extent and direction of variability is improbable on the second measurement in the same individual. Expectedly successive measurements will consequently move closer to the middle 'regress to the mean.' Outcomes from other patients who

initially were nearer to the mean may now fall closer to the extremes of the distribution. This phenomenon can be used to demonstrate treatment efficacy of trials with individuals with high values of a measurement and is regarded as a crucial component in the gold standard of randomized placebo-controlled prospective trials (Bossuyt 2009,). The cause of the variability between two measurements is logically an indication of a legitimate change as opposed to the background noise of quantitative irregularity. Comparably, the smaller the total intra-individual variability, the more likely a particular absolute change is indicated. Lastly, the less probable the observed change caused by variability, the more definite the variations. The least significant change 'LSC' notion embodies these three rudiments (Phillips 2009).

The most common misused tests

Global efforts are in place to improve the allocation of medical resources. Surprisingly the various number of tests available for diagnosing a particular condition actually is detrimental for consistency. This underlines a focus of this paper: the significance of healthcare practitioners utilizing the highest levels of evidence for diagnostic accuracy (White et al 2011, The Joanna Briggs Institute 2015). Nine organizations have collectively developed a top 5 list of assays, treatments, or services that are evidently are needless or require a thorough discussion to make an informed decision about the benefits and risks involved. 'Choosing Wisely' is an initiative developed by the American Board of Internal Medicine Foundation (Bulger et al 2013). A few of the guideline examples include the necessity of ordering an MRI for a new case of low back pain, not ordering an EMG for low back issues unless there is concurrent leg pain, performing an exercise stress test for patients without the signs and symptoms of cardiovascular disease, do not recommend prolonged use of over-the-counter medications for headache (Hudzik et al 2014). These are just a few of the numerous guidelines heavily emphasizing evidence-based clinical practice. This has tremendous potential to reduce redundancy, improve patient outcome and control costs. {http://choosingwisely.org/?page_id=13} for entire list of 45.) As diagnostic and screening assay precision improve, the misuse of testing will of course, simultaneously decrease (Bulger et al 2013).

Future of Laboratory tests

The rate of technological evolution is continuously improving the accuracy of clinical tests, with modern innovative advancements creating another era of laboratory diagnostics (Bossuyt 2003, Bossuyt et al 2007, Walley 2008, Tozzoli 2013). This campaign is driven by the demands for improvements in speed, cost, ease of performance, patient safety and accuracy (Campbell et al 2015). This presents promise for improving contemporary methods to better identify pathological states refining treatment efficiency through earlier and more precise diagnosis thus changing our reliance on clinics and hospitals reducing the burden (Waters 2011, Drucker and Krapfenbauer 2013).

The developing field of human genome is quickly changing the depth of ability in defining RR. It is now understood that genotype and phenotypes of an individual influence analytes. For example, HDL cholesterol concentrations are lower in individuals carrying the Apo A1Milano mutation (Bekaert et al 1993, Boyd 2010). A dominant focus in this field is epigenetics. There has been significant growth in knowledge of the relationship between epigenetic changes influencing neoplasia; this is a groundbreaking clinical benefit as cancer biomarkers. Specific malignancies sensitive to specific cytotoxic chemotherapies may hold potential for forecasting which patients will benefit from newer targeted agents directed at oncogenes. Epigenetic aberrations influence various aspects of tumorigenesis, eventually encouraging the selection of neoplastic cells with increasing pathogenicity. Identifying the associative alterations to predict and improve prognostic biomarkers is an invaluable medical goal. Global analysis strategies are now in place swiftly improving our understanding of the epigenome and promises to boost the identification of epigenomic platforms underlying cancer progression and treatment response (Chan and Baylin 2010). The UK Genetic Testing Network (www.ukgtn.nhs.uk) has appraised over 89 tests, of which 70% were considered acceptable (Walley et al 2008). Melzer and colleagues outline the problems in genetic testing, particularly relating to the evaluation, and have offer methods of overcoming them (Walley et al 2008, Melzer 2008).

Another method with immense potential is Microfluidics. A technology characterized by the engineered manipulation of fluids at the submillimetre scale. Rapid sample processing and the precise control of fluids in an assay, the progress made by lab-on-a-chip microtechnologies in recent years allowing Rapid sample processing to precisely control fluids in an assay (Sackmann, et al 2014).

Approximately 50 % of the early-stage pipeline assets and 30 % of late-stage molecular entities of the pharmaceutical companies comprise the use of specific biomarkers (Chow et al 2013). Additionally, molecular diagnostic tests (also molecular genetic testing or MDx), advance personalized medicine by detecting and measuring proteins, nucleic acids, or metabolites variations, represent the fastest developing segment in the diagnostic (Dx) market enjoying healthy 10 % annual growth and likely to achieve USD 12.78 billion by 2018 (Carson 2014, Akhmetov et al 2015).

Personalized RR as the new gold standard

Reference ranges tend to give the impression of definite thresholds that distinctly separate "healthy" or "unhealthy" values, when in fact there are generally continuously increasing risks with increased distance from usual or optimal values. The boundaries between healthy and pathological states are convoluted areas subjective to many biological factors. This is why the ideal of a single threshold is problematic and debatable: It prevents roadblocks for advancement in an era of medicine focused on substantiation and the precise distinction of existing pathology. It is essential to know how to utilize a continuous biomarker or analytic test. Under the normal conjecture that higher estimations of the biomarker are connected with the illness, this segregated allocation is generally in view of a cut-off worth, such that relying upon whether the individual is determined to be healthy or unhealthy (Whiteley et al 2011, López-Ratón 2016).

Discussion

In synopsis, the points of significance in this paper: there a numerous factors that health care professionals can and must further develop in reference ranges guidelines for diagnosing and screening of disease. Elements from international medical conglomerations are gaining momentum with monumental advancement. Harnessing the immense potential in the modern foundation of evidence-based medicine principles, to refine irrefutable science and eliminate inefficacies, therefore developing the understanding of the nearly inconceivable biological symphony of the human entity and its infinite detectable variations that can possibly be measured. Bringing universal health care to the point where disease detection has expanded to the realm where pathology probabilities are so well understood that a point of revolutionary comprehension can transpire, almost eliminating disease. This will undeniably be a pivotal era in human history.

Conclusion

Laboratory diagnostics is a multifaceted field, consisting of numerous fundamental influences to evidence-based medical practice improving clinical decision-making by use of minimally invasive testing (Plebani 2010, Lippi & Mattiuzzi 2013). The foremost elements of clinical laboratory services are comprised of quality and accuracy of testing, turnaround time, the nature of analysis and additional services provided, expenditures, revenues along with certification or accreditation to reliable standards (Lippi & Mattiuzzi 2013). The incomplete wisdom of reference ranges creates a system of inconsistency and lowers quality of care. This paper identified the deficient validity and the critical need for more reliable medical guidelines. A goal of this review is to shed light and inspire others to push this field. The understanding and growth of this science is vital to the future of healthcare.

References

Adeli, Khosrow, et al. "Complex biological profile of hematologic markers 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* 61.8 (2015): 1075-1086.

Akhmetov, Ildar, and Rostyslav V. Bubnov. "Assessing Value of Innovative Molecular Diagnostic Tests in the Concept of Predictive, Preventive, and Personalized Medicine." *The EPMA Journal* 6 (2015): 19. *PMC*. Web. 13 May 2016.

Altman DG, Bland JM: Diagnostic tests. 1. Sensitivity and specificity. BMJ 1994;308: 1552.

Baloch, Z., et al. "National Academy of Clinical Biochemistry." *Laboratory medicine practice guidelines. Laboratory support for the diagnosis and monitoring of thyroid disease. Thyroid* 13.1 (2003): 3-126.

Beastall, G. H. (2013). Adding value to laboratory medicine: a professional responsibility. *Clinical Chemistry and Laboratory Medicine*, *51*(1), 221-227.

Bossuyt X, Verweire K, Blanckaert N. Laboratory medicine: Challenges and opportunities. *Clin Chem*. 2007;53:1730-1733.

Boyd, James C. "Defining laboratory reference values and decision limits: populations, intervals, and interpretations." *Asian J Androl* 12.1 (2010): 83-90.

Begg CB. Biases in the assessment of diagnostic tests. Stat Med 1987;6:411–23.

Bekaert ED, Alaupovic P, Knight-Gibson CS, Franceschini G, Sirtori CR. Apolipoprotein A-I Milano: sex-related differences in the concentration and composition of apoA-I- and apoB-containing lipoprotein particles. J Lipid Res 1993; 34: 111–23.

Berwick, Donald M., and Andrew D. Hackbarth. "Eliminating waste in US health care." *Jama* 307.14 (2012): 1513-1516.

Biswas, S. S., et al. "Evaluation of Imprecision, Bias and Total Error of Clinical Chemistry Analysers." *Indian Journal of Clinical Biochemistry* 30.1 (2015): 104-108.

Bjoro T, Holmen J, Kruger O, et al. Prevalence of thyroid disease, thyroid dysfunction and thyroid peroxidase antibodies in a large unselected population: the Health Study of Nord-Trondelag (HUNT). *Eur J Endocrinol* 2000; 143: 639–37.

Blankenstein, Marinus A. "Reference intervals–ever met a normal person?." *Annals of Clinical Biochemistry: An international journal of biochemistry and laboratory medicine* 52.1 (2015): 5-6.

Bossuyt PM. Interpreting diagnostic test accuracy studies. Semin Hematol 2008;45:189-195.

Bossuyt, Patrick M., et al (A). "The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration." *Annals of internal medicine* 138.1 (2003): W1-12.

Bossuyt, Patrick M., et al (B). "Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative." *Clinical chemistry and laboratory medicine* 41.1 (2003): 68-73.

Bossuyt, Patrick MM, and Kirsten McCaffery. "Additional patient outcomes and pathways in evaluations of testing." *Medical Decision Making* 29.5 (2009): E30-E38.

Bulger, John, et al. "Choosing wisely in adult hospital medicine: five opportunities for improved healthcare value." *Journal of hospital medicine* 8.9 (2013): 486-492.

Burd, Eileen M. "Validation of laboratory-developed molecular assays for infectious diseases." *Clinical Microbiology Reviews* 23.3 (2010): 550-576.

Campbell, Jared M., et al. "Diagnostic test accuracy: methods for systematic review and meta-analysis." *International journal of evidence-based healthcare* 13.3 (2015): 154-162.

Carson J. Molecular methods rapidly gain ground in infectious disease diagnostics. Frost & Sullivan. 2014.

Carter B. The results are in and they are positive: it's time to optimize the utilization of laboratory tests. CADTH conference presentation [Internet]. St.John's: Pathology and Laboratory Medicine, Province of Newfoundland and Labrador; 2014.

Chandra, Piyanka E., et al. "Novel multiplex technology for diagnostic characterization of rheumatoid arthritis." Arthritis Res Ther 13.3 (2011): R102.

Chow IP, Bryant T, Swanson S, Schoeninger B. Dissecting the value of companion diagnostics. IMS Consulting Group. 2013.

Chan, Timothy A., and Stephen B. Baylin. "Epigenetic biomarkers." *Therapeutic Kinase Inhibitors*. Springer Berlin Heidelberg, 2010. 189-216.

Code of Federal Regulations. 2009. Title 42. Public health, vol. 4, chapter V. Health Care Financing Administration, Department of Health and Human Services, part 493. Laboratory requirements, section 493.1253. Standard: establishment and verification of performance specifications. U.S. Govern- ment Printing Office, Washington, DC.

Dayan, C. M., Saravanan, P., & Bayly, G. (2002). Whose normal thyroid function is better—yours or mine? The Lancet, 360(9330), 353–354. doi:10.1016/S0140-6736(02)09602-2)

Deeks, Jonathan J. "Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests." *British Medical Journal* 323.7305 (2001): 157.

Drucker, Elisabeth, and Kurt Krapfenbauer. "Pitfalls and limitations in translation from biomarker discovery to clinical utility in predictive and personalised medicine." *EPMA J* 4.1 (2013): 7.

European IVD Market Statistics Report. Available at: http://www.edma-ivd.eu/index.php?mact=EuropeanIVD Market Statistics cs,me8bc7,view,1&me8bc7details=26&me8 bc7returnid=104&page=104. [Accessed November 2014]

Eusebi, Paolo. "Diagnostic accuracy measures." *Cerebrovascular Diseases* 36.4 (2013): 267-272.

Forsman RW. Why is the laboratory an afterthought for managed care organizations? Clin Chem. 1996;42(5):813–6. [PubMed]

Fushimi Y, Miki Y, Kikuta K, Okada T, Kanagaki M, Yamamoto A, et al. Comparison of 3.0- and 1.5-T three-dimensional time-of-flight MR angiography in moyamoya disease: preliminary experience. Radiology [Internet]. 2006 Apr [cited 2012 Aug 24];239(1):232-7. Available from: http://radiology.rsna.org/content/239/1/232.full.pdf+html

Grasbeck, R., and N. E. Saris. "Establishment and use of normal values." *Scandinavian Journal of Clinical & Laboratory Investigation*. PO BOX 2959 TOYEN, JOURNAL DIVISION CUSTOMER SERVICE, N-0608 OSLO, NORWAY: SCANDINAVIAN UNIVERSITY PRESS, 1969.

Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. CMAJ 1986;134:587–94.

Hak AE, Pols HAP, Visser TJ, Drexhage HA, Hofman A, Witteman JCM. Subclinical hypothyroidism is an independent risk factor for atherosclerosis and myocardial infarction in elderly women: the Rotterdam study. *Ann Intern Med* 2000; 132: 270–78.

Hallworth M. Demonstrating the impact of laboratory medicine on clinical outcomes. Clin Chem Lab Med [Internet]. 2014 Jun [cited 2014 Aug 20];52(Supplement):s34. Symposium abstracts. Available from: http://www.degruyter.com/view/j/cclm.2014.52.issue-s1/issue-files/cclm.2014.52.issue-s1.xml (Presented at IFCC WorldLab Istanbul 2014 - Istanbul, 22-26 June 2014).

Hanson, Curtis, and Elizabeth Plumhoff. "Test Utilization and the Clinical Laboratory." *Canadian Journal of Pathology* 4.4 (2012).

Hauser, Ronald G., and Brian H. Shirts. "Do We Now Know What Inappropriate Laboratory Utilization Is?." *American Journal of Clinical Pathology* 141.6 (2014): 774-783.

Hickner, John, et al. "Primary care physicians' challenges in ordering clinical laboratory tests and interpreting results." *The Journal of the American Board of Family Medicine* 27.2 (2014): 268-274.

Hogg W, Baskerville N, Lemelin J. Cost savings associated with improving appropriate and reducing inappropriate preventive care: cost-consequences analysis. *BMC Health Serv Res.* 2005;5:20.

Horowitz, Gary L., Sousan Altaie, and James C. Boyd. *Defining, establishing, and verifying reference intervals in the clinical laboratory; approved guideline*. CLSI, 2010.

Hudzik, Bartosz, Michal Hudzik, and Lech Polonski. "Choosing Wisely: Avoiding Too Much Medicine." *Canadian Family Physician* 60.10 (2014): 873–876. Print.

Huma, Tanzeel, and Usman Waheed. "THE NEED TO ESTABLISH REFERENCE RANGES." J Pub Health Bio Sci. 2013;2(2):188-190. Journal of Public Health and Biological Sciences Vol. 2, No. 2 Apr – Jun 2013, p.188-190 ISSN 2305-8668 (Print) 2307-0625 (Online) URL: http://www.jphbs.com

Hunink MG, Krestin GP. Study design for concurrent development, assessment, and implementation of new diagnostic imaging technol- ogy. Radiology 2002;222:604e14.

Ichihara K, Itoh Y, Lam CW, Poon PM, Kim JH, *et al*. Sources of variation for commonly measured serum analytes among six Asian cities and consideration of common reference intervals. Clin Chem 2008; 54: 356–65.

Ichihara K, Itoh Y, Min WK, Sook FY, Lam CW, *et al*. Diagnostic and epidemiological implications of regional differences in serum concentrations of proteins observed in six Asian cities. Clin Chem Lab Med 2004; 42: 800–9.

Irwig L, Glasziou P, Wilson A, Macaskill P. Estimating an individual's true cholesterol level and response to intervention. JAMA 1991;266:1678-85.

Jaeschke R, Guyatt G, Sackett DL. 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 1994;271:389–91.

Jaeschke R, Guyatt GH, Sackett DL. 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 1994;271:703–7.
JAMA 1984;252:2418–22.

Jones, Graham, and Antony Barker. "Reference intervals." *Clin Biochem Rev* 29.Suppl 1 (2008): S93-7.

Kanchanaraksa, Sukon. "Evaluation of diagnostic and screening tests: validity and reliability." *Baltimore: John Hopkins University* (2008).

Katayev, Alex, Claudiu Balciza, and David W. Seccombe. "Establishing reference intervals for clinical laboratory test results." *American Journal of Clinical Pathology* 133.2 (2010): 180-186.

Kim Futrell, M. T. "Meaningful Medical Analytics: Driven by Laboratory Data Integration." (2015).

Korevaar, Daniël A., et al. "Literature survey of high-impact journals revealed reporting weaknesses in abstracts of diagnostic accuracy studies." *Journal of clinical epidemiology* 68.6 (2015): 708-715.

Leeflang, M. M., Deeks, J. J., Takwoingi, Y., & Macaskill, P. (2013). Cochrane diagnostic test accuracy reviews. *Systematic reviews*, *2*(1),

Lippi et al.: Patient misidentification and laboratory medicine. Clin Chem Lab Med 2009;47(2):143–153 2009 by Walter de Gruyter • Berlin • New York. DOI 10.1515/CCLM.2009.045

Lippi, G., & Mattiuzzi, C. (2013). Testing volume is not synonymous of cost, value and efficacy in laboratory diagnostics. *Clinical Chemistry and Laboratory Medicine*, *51*(2), 243-245.

Plebani, Mario. "Laboratory diagnostics in the third millennium: where, how and why. Foreword." *Clinical chemistry and laboratory medicine: CCLM/FESCC* 48.7 (2010): 901-902.

López-Ratón, Mónica, et al. "Confidence intervals for the symmetry point: an optimal cutpoint in continuous diagnostic tests." *Pharmaceutical statistics* (2016).

Mindemark, Mirja, and Anders Larsson. "Longitudinal trends in laboratory test utilization at a large tertiary care university hospital in Sweden." *Upsala journal of medical sciences* 116.1 (2011): 34-38. Forsman, Rodney W. "Why is the laboratory an afterthought for managed care organizations?." *Clinical Chemistry* 42.5 (1996): 813-816.

Ochodo, E. A., & Leeflang, M. M. (2012). Systematic reviews of diagnostic test accuracy for evidence-based diagnostic practice in Africa. *African Journal of Laboratory Medicine*, *1*(1), 3-pages.

LaboratorySM, Great Smokies Diagnostic. "The Functional Physiologic Range: The Guide to Assessing Optimal Health." (2001).

Lalkhen AG, McCluskey A. Clinical tests: sensitivity and specificity. Continuing Education in Anaesthesia, Critical Care & Pain. 2008;8(6):221-3

Lalkhen, Abdul Ghaaliq, and Anthony McCluskey. "Clinical tests: sensitivity and specificity." *Continuing Education in Anaesthesia, Critical Care & Pain* 8.6 (2008): 221-223.

Leeflang, Mariska MG, et al. "Variation of a test's sensitivity and specificity with disease prevalence." *Canadian Medical Association Journal* 185.11 (2013): E537-E544.

Lijmer, Jeroen G., Mariska Leeflang, and Patrick MM Bossuyt. "Proposals for a phased evaluation of medical tests." *Medical Decision Making* 29.5 (2009): E13-E21.

Linardou H, Dahabreh J, Kanaloupiti D, et al. Assessment of somatic *k-RAS* mutations as a mechanism associated with resistance to EGFR-targeted agents: a systematic review and meta-analysis of studies in advanced non-small-cell lung cancer and metastatic colorectal cancer. The Lancet Oncology 2008;9(10):962-72.

Liu Z, Abdullah A, Baskin L, Lewis G, Kelter G, Naugler C. An intervention to reduce laboratory utilization of referred-out tests. Lab Med. 2012;43(5):164-67.

Lord SJ, Irwig L, Bossuyt PM. Using the principles of randomized controlled trial design to guide test evaluation. Med Decis Making 2009;29:E1-E12.

Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? Ann Intern Med 2006;144:850-855.

M.E. Porter, E.O. Teisberg, Redefining Health Care: Creating Value-based Competition on Results, xviiHarvard Business School Press, Boston, Massachusetts, 2006. (506 pages).

Mariska M.G. Leeflang PhD, Anne W.S. Rutjes PhD, Johannes B. Reitsma MD PhD, Lotty Hooft PhD, Patrick M.M. Bossuyt PhD. CMAJ, August 6, 2013, 185(11) E537

Melzer D, Hogarth S, Liddell K, Ling T, Sanderson S, Zimmern RL. Genetic tests for common diseases: new insights, old concerns. *BMJ* 2008 doi: 10.1136/bmj.39506.601053.BE.

Metz CE. Basic principles of ROC analysis. Semin Nucl Med. 1978;8(4):283-98.

Michalopoulou G, Alevizaki M, Piperingos G, et al. High serum cholesterol levels in persons with "high-normal" TSH levels: should one extend the definition of subclinical hypothyroidism? *Eur J Endocrinol* 1998; 138: 141–45.

Miyakis S, Karamanof G, Liontos M, Mountokalakis TD. Factors contributing to inappropriate ordering of tests in an academic medical department and the effect of an educational feedback strategy. Postgrad Med J. 2006 Dec;82(974):823-9.

Morehouse P. The results are in and they are positive: it's time to optimize the utilization of laboratory tests [Internet]. Sydney (NS): Cape Breton District Health Authority; 2013. [cited 2014 Mar 20]. Available from: http://labtest.cadth.ca/media/labtest- resources/Moorehouse.pdf

Mower WR. Evaluating bias and variability in diagnostic test reports. Ann Emerg Med 1999;33:85–91.

National Committee for Clinical Laboratory Standards. How to Define and Determine Reference Intervals in the Clinical Laboratory: Approved Guideline. NCCLS Document C28A: Villanova, PA; NCCLS, 1995:28.

Naugler C. A perspective on laboratory utilization management from Canada. Clin Chim Acta. 2014;427:142-4.

Newman-Toker, David E., and Peter J. Pronovost. "Diagnostic errors—the next frontier for patient safety." *JAMA* 301.10 (2009): 1060-1062.

O'Donnell, Michael P. "Publisher's Notes." *American journal of health promotion* 2.1 (1987): 4-4.

Okeh, U. M., and C. N. Okoro. "Evaluating measures of indicators of diagnostic test performance: fundamental meanings and formulars." *Journal of Biometrics & Biostatistics* 2012 (2012).

Panagiotakos, Demosthenes B. "The value of p-value in biomedical research." *The open cardiovascular medicine journal* 2.1 (2008).

Panteghini M. The future of laboratory medicine: Understanding the new pressures. Clin Biochem Rev. 2004;25:207–215.

Phillips, Pat. "Pitfalls in interpreting laboratory results." *Aust Prescr* 32 (2009): 43-46.
Price, Christopher P. "Evidence-based laboratory medicine: is it working in practice." *Clin Biochem Rev* 33 (2012): 13-19.

Rector, Thomas S., Brent C. Taylor, and Timothy J. Wilt. "Systematic review of prognostic tests." *Journal of general internal medicine* 27.1 (2012): 94-101.

Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. JAMA 1995;274:645–51.

Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol* 2005;58:982–90.

Reitsma, Johannes B., et al. "Systematic reviews of studies quantifying the accuracy of diagnostic tests and markers." *Clinical chemistry* 58.11 (2012): 1534-1545.

Ricos C, Perich C, Minchinela J, Álvarez V, Simón M, Biosca C, et al. Application of biological variation: a review. Biochem Med. 2009;19(3):250–259. doi: 10.11613/BM.2009.023

Rohr, Ulrich-Peter, et al. "The value of in vitro diagnostic testing in medical practice: a status report." *PloS one* 11.3 (2016): e0149856.

Rollins G. The path to better test utilization: why labs should step-up physician education, consultation. Clin Lab News [Internet]. 2012 [cited 2014 Mar 20];38(9). Available from: http://www.aacc.org/publications/cln/2012/September/Pages/TestUtilization.aspx#

Sackett, D. L., and R. B. Haynes. "The architecture of diagnostic research." *British Medical Journal* 324.7336 (2002): 539.

Sackett, David L., et al. "Evidence based medicine: what it is and what it isn't." *Bmj* 312.7023 (1996): 71-72.

Sackmann, Eric K., Anna L. Fulton, and David J. Beebe. "The present and future role of microfluidics in biomedical research." *Nature* 507.7491 (2014): 181-189.

Sartore-Bianchi A, Martini M, Molinari F, et al. *PIK3CA* mutations in colorectal cancer are associated with clinical resistance to EGFR-targeted monoclonal antibodies. Cancer Res 2009;69(5):1851-7.

Sheps SB, Schechter MT. The assessment of diagnostic tests. A survey of current medical research.

Siest, Gerard, et al. "The theory of reference values: an unfinished symphony." *Clinical Chemistry and Laboratory Medicine* 51.1 (2013): 47-64.

Šimundić, Ana-Maria. "Measures of diagnostic accuracy: basic definitions." *Med Biol Sci* 22.4 (2008): 61-5.

St John A, Price CP. Economic Evidence and Point-of-Care Testing. *The Clinical Biochemist Reviews*. 2013;34(2):61-74.

St John, A., Edwards, G., Fisher, S., Badrick, T., Callahan, J., & Crothers, J. (2015). A call for a value based approach to laboratory medicine funding. *Clinical biochemistry*, *48*(13), 823-826.

The Joanna Briggs Institute, Joanna Briggs Institute Reviewers' Manual: 2015 edition / Supplement, The systematic review of studies of diagnostic test accuracy

Thienpont, Linda M., et al. "Determination of free thyroid hormones." *Best Practice & Research Clinical Endocrinology & Metabolism* 27.5 (2013): 689-700.

Tozzoli, R., Bonaguri, C., Melegari, A., Antico, A., Bassetti, D., & Bizzaro, N. (2013). Current state of diagnostic technologies in the autoimmunology laboratory. *Clinical Chemistry and Laboratory Medicine*, *51*(1), 129-138.

Ups J Med Sci. 2011 February; 116(1): 34–38. Published online 2011 February 11. doi: 10.3109/03009734.2010.528071

Utah Governor Huntsman recognizes medical laboratory week. *Clinical Lab Products*. Issue Stories. May 2006. http://www.clpmag.com/issues/articles/2006-05_03.asp. Accessed January 1, 2009.

van Walraven C, Raymond M. Population-based study of repeat laboratory testing. Clin Chem [Internet]. 2003 Dec [cited 2014 Mar 20];49(12):1997-2005. Available from: http://www.clinchem.org/content/49/12/1997.full.pdf+html

Verbrugghe S. Initiatives to optimize the utilization of laboratory tests. Ottawa: Canadian Agency for Drugs and Technologies in Health; 2014. (Environmental Scan; issue 44).

Walley, T. (2008). Evaluating laboratory diagnostic tests. *BMJ*, *336*(7644), 569-570.

Walter LC, Fung KZ, Kirby KA, et al. Five-year downstream outcomes following prostate-specific antigen screening in older men. *JAMA Intern Med.* 2013;173:866-873.

Walton RM. Establishing reference intervals: health asa relative concept. Semin Avian Exotic Pet Ped. 2001;10:67–71.

Waters HR, Korn RJ Jr, Colantuoni E, et al. The business case for quality: economic analysis of the Michigan Keystone Patient Safety Program in ICUs. *Am J Med Qual*. 2011;26(5):333-339.

Wayne, P. "Defining, establishing, and verifying reference intervals in the clinical laboratory: approved guidelinethird edition." *CLSI documalet C28-A3c. 3rd ed. Wayne (PA): Clinical and Laboratory Standards Institute* (2008).

Wener, M. H. (2011). Multiplex, megaplex, index, and complex: the present and future of laboratory diagnostics in rheumatology. *Arthritis research & therapy*, *13*(6), 1-3.

White S, Schultz, T., Enuameh, YAK., ed. Synthesizing evidence of diagnostic accuracy. Philadelphia, USA: Lippincott Williams and Williams 2011.

Whiteley W, Wardlaw J, Dennis M, Lowe G, Rumley A, Sattar N, Welsh P, Green A, An- drews M, Graham C, Sandercock P: Blood biomarkers for the diagnosis of acute cerebro- vascular diseases: a prospective cohort study. Cerebrovasc Dis 2011;32:141–147.

Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol. 2003; 3:25.

Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Ann Intern Med* 2004;140:189–202.

Whiting, Penny F., et al. "QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies." *Annals of internal medicine* 155.8 (2011): 529-536.

Willis BH, Quigley M. Uptake of newer methodological developments and the deployment of meta-analysis in diagnostic test research: a systematic review. BMC Med Res Methodol. 2011; 11:27.

Wolcott J, Schwartz A, Goodman C. Laboratory medicine: a national status report [Internet]. Falls Church (VA): The Lewin Group; 2008. Available from: https://www.futurelabmedicine.org/pdfs/2007%20status%20report%20laboratory_medicine_-_a_national_status_report_from_the_lewin_group.pdf

Yadav, Dharmveer, et al. "Reference Interval for Renal Profile parameters in North Indian Population from Rajasthan." *International Journal of Advances in Scientific Research* 1.5 (2015): 233-238.http://articles.chicagotribune.com/2011-11-21/a-z/sc-health-1123-bloodwork-20111121_1_labs-range-glucose.

Youden WJ (1950) Index for rating diagnostic tests. Cancer 3: 32-35.

Zidan, Marwan, Ronald L. Thomas, and Thomas L. Slovis. "What you need to know about statistics, part II: reliability of diagnostic and screening tests." *Pediatric radiology* 45.3 (2015): 317-328

Zou KH, O'Malley AJ, Mauri L. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. Circulation. 2007;115(5):654-7.