Asian Journal of Andrology (2010) 12: 83–90

www.nature.com/aja

Perspectives

Defining laboratory reference values and decision limits: populations, intervals, and interpretations

James C. Boyd

Department of Pathology, University of Virginia Health System, Charlottesville, VA 22908-0168, USA

Abstract

This article provides a brief overview of various approaches that may be utilized for the analysis of human semen test results. Reference intervals are the most widely used tool for the interpretation of clinical laboratory results. Reference interval development has classically relied on concepts elaborated by the International Federation of Clinical Chemistry Expert Panel on Reference Values during the 1980s. These guidelines involve obtaining and classifying samples from a healthy population of at least 120 individuals and then identifying the outermost 5% of observations to use in defining limits for two-sided or one-sided reference intervals. More recently, decision limits based on epidemiological outcome analysis have also been introduced to aid in test interpretation. The reference population must be carefully defined on the basis of the intended clinical use of the underlying test. To determine appropriate reference intervals for use in male fertility assessment, a reference population of men with documented time to pregnancy of < 12 months would be most suitable. However, for epidemiological assessment of semen testing results, a reference population made up of unselected healthy men would be preferred. Although reference and decision limits derived for individual semen analysis test results will undoubtedly be the interpretational tools of choice in the near future, in the long term, multivariate methods for the interpretation of semen analysis alone or in combination with information from the female partner seem to represent better means for assessing the likelihood of achieving a successful pregnancy in a subfertile couple.

Asian Journal of Andrology (2010) 12: 83–90. doi: 10.1038/aja.2009.9

Keywords: decision limits, fertility assessment, human semen testing, likelihood ratios, reference values, semen analysis

1 Introduction

This issue highlights the publication of the fifth edition of the World Health Organization (WHO) Laboratory Manual for the Examination and Processing of Human Semen [1]. Results will soon become available for an in-depth study of over 4 500 individuals undertaken by the Editorial Committee to establish reference values for use in human semen analysis [2]. These two documents contain the latest and the most comprehensive available information on techniques for assessing the characteristics of human semen and the likelihood of male fertility. Although the data gathered from human semen analysis comprise only a portion of the information used for assessing the likelihood of achieving viable pregnancy, careful interpretation of these results can help guide the selection of further examinations. The availability of extensive information regarding the setting of reference values for human semen tests will allow these tests to be better evaluated using established metrics for traditional clinical laboratory tests. The intent of this article is to examine the utility of reference values compared with possible alternative approaches for the evaluation of semen testing results, including decision limits and probability-based methods such as likelihood ratios.

Measurements are frequently interpreted by comparison with earlier measurements taken from the population at large. In clinical medicine, special attention is given to assem-

Correspondence to: Dr James C. Boyd, Department of Pathology, University of Virginia Health System, P.O. Box 800168, Charlottesville, VA 22908-0168, USA.
Fax: +1-434-243-5930 E-mail: jboyd@virginia.edu
Received: 1 January 2009 Revised: 15 January 2009
Accepted: 19 January 2009

bling sets of measurements in carefully defined groups of individuals that can provide a frame of reference for the clinical interpretation of subsequent laboratory test measures. These sets of 'reference' values or their predecessors, 'normal' values, have been used by physicians for generations to assist them in interpreting biochemical or physiological measurements in patients [3, 4]. Advanced theory on statistical approaches for the determination of reference values is available in the books by Harris and Boyd [5] and Horn and Pesce [6]. In addition, the Clinical Laboratory Standards Institute and the International Federation of Clinical Chemistry (IFCC) have issued detailed sets of guidelines regarding the determination of reference limits [7–13]. Therefore, only a brief overview of these concepts is provided here.

Decision limits are another tool that may aid in the interpretation of test results in specific clinical circumstances. In recent years, the distinction between reference limits and decision limits has become blurred by the introduction of limits that are based on epidemiological evidence regarding likely clinical outcomes for patients who have laboratory test results above or below the 'reference' limits. An example of such blurring is the current set of upper reference limits for cholesterol recommended by an expert panel for the National Cholesterol Education Program (NCEP) [14], which is further discussed below. To clarify the distinction between classical reference limits and decision limits, we must delve into the definition of normality, how reference intervals are defined, and factors that can influence reference intervals.

2 What is normal?

Clinical laboratories have for many years reported results relative to 'normal' values. However, as pointed out by the philosopher Edmund Murphy [15], the word 'normal' can be interpreted from many frames of reference. Using some of Murphy's definitions, a 'normal' sperm concentration might be: (1) the most representative sperm concentration as defined by the mean; (2) the most commonly encountered sperm concentrations as defined by an interval (i.e., the usual laboratory reference interval); (3) sperm concentrations associated with fertility; (4) a committee's consensus of 'approved' sperm concentrations; or (5) the ideal sperm concentration. Physicians are usually interested in 'normality' in terms of definitions (2) or (3). Typically, *reference values* or *reference intervals* are established for each laboratory test to delineate the range of values that would usually be encountered in a 'healthy' population.

Normality is relative. The 'normal' sperm concentration, for example, is influenced by many non-disease-related factors including age, endocrine status, physical activity, duration of abstinence, volume of ejaculate, and history of fertility. Cooper *et al*. [2] found that men in a population group with time to pregnancy of ≤ 12 months have demonstrably higher sperm concentrations than men in an unscreened population or men who have gone through screening examinations. Thus, reference values must be defined in terms of a specific reference population. In this case, if the desired comparison group consists of men with demonstrated fertility and time to pregnancy of ≤ 12 months, reference values should be defined for that specific population group and may differ from those defined for the population at large. Many have therefore replaced the label 'normal' interval with 'reference' interval, as these values merely provide a frame of reference for interpreting other results.

3 Definition of reference interval

The reference interval for many laboratory tests is defined by threshold values between which the test results of a specified percentage (usually 95%) of apparently healthy individuals would fall. The threshold or limiting values for the reference interval are usually the 0.025 and 0.975 fractiles of the test result distribution in the reference population (Figure 1). This definition results in exclusion of the 2.5% of individuals with the lowest results and the 2.5% of individuals with the highest results from the reference interval. In case the clinical interest is only in 'low' results and high test results are not indicative of pathology (or, conversely, if only high test results are of interest), one-sided reference intervals are defined that exclude

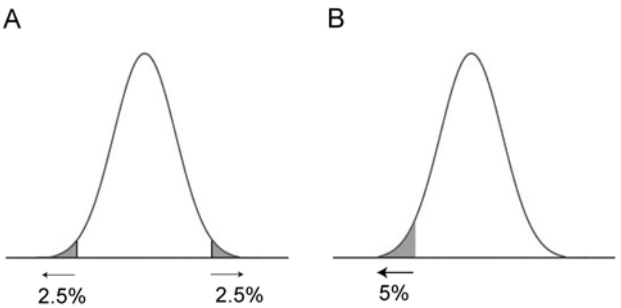

Figure 1. The 95% reference intervals are derived by identifying the most outlying 5% of observed values in a reference population. Most often, these outlying observations are split evenly between the ends of the test result distribution in the reference population, 2.5% at each end of the distribution, resulting in a two-sided reference interval (A). For some tests (e.g., sperm concentration), only low values are of clinical concern, and therefore the 5% of outlying observations for the reference interval are identified and excluded only from the low end of the distribution, resulting in a one-sided reference interval (B).

only the 5% of the population in the 'abnormal' tail of the distribution. In interpreting sperm concentration, for instance, only low concentrations are likely to be of clinical concern, and the use of the 0.05 fractile of the reference population as a one-sided lower reference limit makes the most sense.

Although this statistical definition of the reference interval has been applied to the majority of laboratory tests, reference intervals have more recently been defined on the basis of analysis of clinical outcomes. The reference limits for cholesterol defined by the NCEP [14] provide an example of this approach. Whereas the 97.5 percentile for cholesterol concentration in the general population lies between 280 and 300 mg dL^{-1} (7.25–7.77 mmol L^{-1}), the upper reference limits for cholesterol as defined by the NCEP are 200 mg dL^{-1} (5.18 mmol L^{-1}), corresponding to approximately the 50th percentile of the population, and 240 mg dL^{-1} (6.22 mmol L^{-1}), corresponding to approximately the 75th percentile. These values were chosen by the NCEP expert panel because they were associated with moderate and high risks for the development of cardiovascular disease in epidemiological outcome studies. For male fertility assessment, it seems reasonable to suggest that defining reference limits for human semen tests based on achieved rates of pregnancy in the reference population may ultimately be the preferred approach.

Another approach for defining reference intervals is emerging from the rapidly expanding knowledge base of the human genome. In this case, it is now known that the most commonly encountered values (reference values) for some analytes vary with the genotype (and/or phenotype) of the individual. Examples of this include haptoglobin (in which concentrations progressively decrease across the Hp 1-1, Hp 2-1, and Hp 2-2 phenotypes) [16] and HDL cholesterol (in which concentrations are lower in individuals carrying the Apo $A1_{Milano}$ mutation) [17]. Several genetic markers are known to have a role in male infertility [18], and it is possible that as yet undefined genetic markers may influence observed reference intervals for tests of human semen or male fertility.

4 Factors in the determination of 'normal' or reference intervals

4.1 Choice of population to study

Many factors must be considered in the determination of normal values or reference intervals. One extremely important factor is the choice of which population to study. Where the range of test values seen in healthy individuals is the primary concern, volunteers should be selected who reflect the overall healthy population. Possible approaches that can be used include studying a 'random' sample from a 'normal' population such as volunteer blood donors, door-to-door contacts, medical students, or medical technologists. Regardless of the reference population selected for study, there is always the potential that the specific group of individuals selected may not be representative of that population. As the same factors that lead individuals to volunteer for such a study (e.g., participants may volunteer owing to an underlying concern about their health, and study organizers have offered the inducement of free laboratory test results or a free medical evaluation) may also have an effect on their test results, the resulting population values may be biased.

If a particular population characteristic guides the definition of the reference population (e.g., demonstrably fertile males with time to pregnancy ≤ 12 months), then the reference population should reflect a random sampling of such individuals. On the other hand, if fertility is not the underlying concern, but rather the epidemiological relationship of an individual's semen analysis results to the population at large, then the most appropriate reference population will be made up of randomly selected healthy men from the general population. The fact that Cooper *et al*. [2] have shown statistically significant differences in semen analysis results between the fertile male population and the healthy male population at large strengthens the argument that where male fertility is the clinically important criterion, only reference values derived from a population of known fertile males should be used. However, the need to collect appropriate documentation of fertility and time to pregnancy for each reference interval makes studying this population a more difficult and expensive proposition than studying volunteers from the population at large. In addition, because of the factors pointed out above, it may be very difficult to recruit a set of volunteers who truly represent a random, unbiased sampling of the population of interest.

It should be pointed out that laboratory test results drawn from healthy populations in different geographical regions exhibit significant variation [19]. Ichihara *et al.* [20] conducted a follow-up study to explore possible causes of the between-city differences. Their study not only confirmed the presence of large between-city differences in several analytes, but also indicated a biological basis for this variation. In both studies, investigators removed the between-laboratory component of variability by analyzing deep-frozen specimens collected from six Asian cities in a single central laboratory. Although such differences may seem surprising, the fact that they were confirmed in two independent studies is compelling and suggests caution regarding the development of 'universal' reference intervals for worldwide application. Thus, although adherence to the fifth edition of the WHO manual for semen analysis [1] will serve to reduce variability between laboratories, the likelihood of regional differences in semen analysis results

that have an underlying biological basis cannot be ignored in future epidemiological investigations that involve semen analysis.

Some researchers have attempted to derive a 'normal' population from easily accessed hospitalized patients by applying selection criteria (e.g., only patients hospitalized for senile cataract removal, cosmetic surgery, herniorrhaphy, or hemorrhoidectomy). Unfortunately, hospitalized patients are likely to have other conditions, diagnosed or undiagnosed, that do not reflect the population at large. All patients used in such studies should be medically screened to ensure that they do not have other ailments that could affect results of the test for which a reference interval is desired.

4.2 Pre-analytical variables

In addition, variables of the study population itself that can affect test results must be carefully assessed and controlled (pre-analytical variables). For various laboratory tests, these variables can include age, diet, sex, circadian rhythm, race, posture, medications, physical activity, socioeconomic status, medical history, and fasting status. For semen analysis, in particular, variables likely to affect results include the age and endocrine status of the reference individual, as well as the period of abstinence before obtaining the semen sample.

4.3 Analytical variation

Clearly, differences in how the test procedure is performed or differences in interpretive criteria can have a major impact on the test results of reference intervals. These factors call for the highest degrees of test standardization and quality assessment. For the analysis of human semen, the fifth edition of the WHO Manual for the Examination and Processing of Human Semen [1] serves as an excellent resource to promote the standardization of test procedures and gives recommendations regarding ongoing quality control.

4.4 Calculation of the reference interval

Once a series of results have been generated for the selected reference population, the reference intervals can be calculated. Two methods have commonly been used for calculation of the reference interval from study data: With the parametric approach, the central 95% boundaries are specified by the mean ± 2SD, if the data follow a Gaussian (normal) distribution or can be transformed to a normal distribution by one of several two-step transformation methods [21] and with the non-parametric approach, the central 95% boundaries are determined by trimming off the lowest and highest 2.5% of observations. The latter method is used for skewed and other non-normal data distributions. The IFCC committee has developed a computer program called REFVAL that implements both parametric and non-parametric methods, including bootstrapping methods for generating confidence intervals around the reference interval limits [22]. Horn *et al.* [23] have described a robust method that provides a non-parametric approach for calculation of reference interval limits that allows use of the smaller reference populations and is more resistant to the effects of outlier results than either the parametric or non-parametric methods described above. The Clinical and Laboratory Standards Institute approved guidelines for determination of reference intervals provide more detailed guidance regarding the application of various calculation methods [7].

As mentioned earlier, outcomes analysis is now being used for the determination of reference values for certain laboratory analytes. In the case of cholesterol, the incidence of cardiovascular disease in cohorts of patients stratified across cholesterol concentrations was utilized by the NCEP in defining risk thresholds for cholesterol. With the data gathered by Cooper *et al*. [2], it may be possible to use a similar approach for setting reference limits based on the relative fertility of individuals whose semen test was at or below given threshold values.

5 Reference intervals and test interpretation

Reference intervals have several advantages in routine clinical applications, including their simplicity, ease of storage and retrieval from laboratory computer systems and pocket notebooks, and their high degree of acceptance by the medical community through long use. However, reference intervals are not ideal for interpreting laboratory results in many circumstances.

Reference intervals have several disadvantages. If the reference interval has been derived from a population dissimilar to the individual tested, it may give a misleading impression of the status of the individual patient. In addition, reference intervals are relatively inflexible instruments and do not take into account any special history or other characteristics of the patient. Thus, if the patient is a strict vegetarian or a diabetic, for example, and either of these conditions affects the results of the laboratory test under consideration, the reference interval for that test as derived from the general healthy population (most of whom are not vegetarians or diabetics) will not provide the correct basis for comparison. Furthermore, the statistical definition of the reference interval may not allow certain clinical uses. As a specific example, because 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.

Owing to the statistical manner in which the 95% reference interval is defined, 5% of normal subjects will be

'abnormal' (i.e., have values for a single test that fall outside of the reference interval). This definition often leads to the misconception that 95% of the diseased individuals will have test results that lie outside the reference interval. This is rarely, if ever, true. Instead, the number of diseased individuals who fall outside the reference interval must be determined by study of the distribution of results in a defined population with the target condition. If we refer to hypothetical distributions of sperm concentration test results in fertile and infertile men (Figure 2), we note that the test results for the fertile and infertile populations show quite a bit of overlap. This occurrence is the rule, not the exception. Many individuals in the infertile population can have results that are within the fertile population reference interval. Thus, just as finding that an individual's result is outside the reference interval does not imply that the man is infertile (because 5% of fertile individuals, by definition, have results in the 'abnormal' range); finding that an individual's result lies within the reference interval does not imply with certainty that this individual will be fertile.

Part of the reason that overlap has been observed in the distributions of test results in fertile and infertile individuals is that determination of fertility is dependent on many variables beyond the tests performed in the examination of semen. Examination of any single test result from a semen examination will not necessarily provide a definitive projection of infertility in a given patient. A potential solution to this problem is to develop a multidimensional reference region, a topic that is too complex to cover here [24]. However, there are other multivariate approaches that are superior to multidimensional reference regions, and these are covered below.

When test results for both fertile and infertile individuals are available, various approaches can be utilized to set *decision limits* for laboratory tests by examining the test *sensitivity* (rate of positive test results in infertile individuals) and test *specificity* (rate of negative test results in fertile individuals) at various test threshold settings. Such thresholds are best set by the use of receiver operating characteristic analysis [25]. Examples of studies that have used this approach for setting decision limits of tests in semen analysis include those by Gunalp *et al*. [26] and Nasr-Esfahani *et al.* [27].

A final disadvantage of the reference interval is that information in a laboratory result is lost when it is converted to 'low', 'normal', or 'high'. Thus, for example, a patient with a sperm concentration of 10 million mL^{-1} will be regarded as having a value below the reference limit, as would a patient whose concentration is < 100 000 mL^{-1}. However, the latter patient is likely to have a completely different clinical evaluation for infertility compared with the former patient. A tool for aiding the interpretation of a laboratory test that retains quantitative information would be a useful addition and we will turn to this topic next.

6 Alternatives to the reference interval

From the above list of disadvantages, we can see that reference intervals are imperfect data interpretational aids. How might we reduce the rigidity and arbitrariness of '95% reference intervals'? One possible solution would be to use a universal scale for reporting results.

6.1 Universal scales

Several universal scales have been suggested. First, results could be reported in terms of the number of stan-

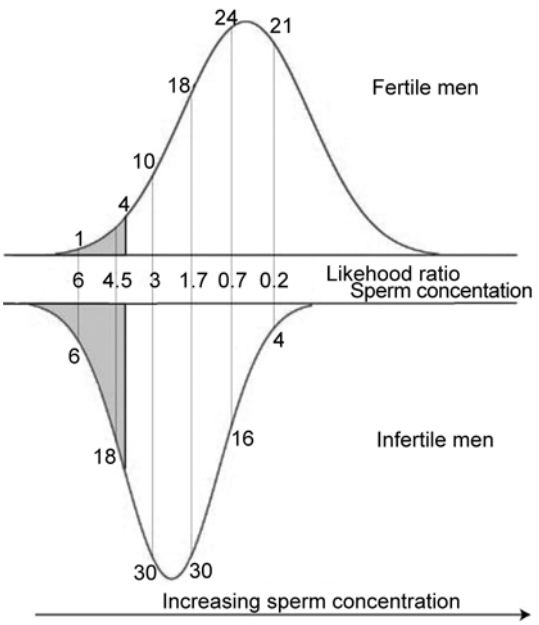

Figure 2. Distributions of test results for sperm concentration are plotted for the hypothetical population subgroups of fertile and infertile men. As these are hypothetical distributions, no specific sperm concentrations are indicated on this graph. Infertile men as a group are displayed as having lower average sperm concentrations than fertile men, and more infertile than fertile men have sperm concentrations below the lower reference limit (shaded area). The relative heights of the distribution curve for each group of men are displayed for several vertical line segments, each of which represents a different sperm concentration. The ratios of distribution heights for the infertile men/fertile men represent the likelihood ratios for infertility at each sperm concentration. Thus, for the line segment at the farthest left, corresponding to a low sperm concentration, the ratio of line heights is 6/1, yielding a likelihood ratio for infertility of 6. Conversely, at the line segment to the farthest right, corresponding to a higher sperm concentration, the ratio of line heights is 4/21, which gives a likelihood ratio for infertility approximately equal to 0.2.

dard deviations away from the population mean. However, for consistency of interpretation, this approach requires that data follow a normal distribution and it does not directly indicate the probability of a fertile individual having a given result.

Another approach would be to report results in percentiles. Although this approach does not require data to follow a normal distribution, it does require that all data from the reference group be available or that a suitable equation be available for transforming a result into a percentile. Computer transformation of results directly to percentiles can be accomplished in two ways: (1) transforming results to a Gaussian scale [20] and then estimating the percentiles or (2) tabular storage of all results from the reference population and matching the patient result with the appropriate reference percentile.

6.2 Likelihood ratios

The goal of a study to evaluate the diagnostic accuracy of a given laboratory test is to establish whether a relationship exists between the test result and a clinically defined end point (e.g., presence or absence of a disease or likelihood of fertility). The strength of any relationship must be evaluated to establish the utility of the marker in predicting disease or outcome in the individual patient. It is assumed that the clinical outcome category (disease *vs.* non-disease, fertile or infertile) can be accurately assessed by some independent criterion (a so-called reference standard or 'gold standard'). In the case of male infertility, definition of such a gold standard may be difficult, as infertility is a multifactorial trait that requires consideration of factors pertaining to both the man and the woman in a given couple.

The *likelihood ratio* of a test value (x) for analyte X is written $L(X = x)$, or simply $L(x)$, and is defined as [28]:

$$L(x) = \frac{\text{Prob}(X = x : \text{Condition})}{\text{Prob}(X = x : \text{No condition})}$$

Likelihood ratios state how many times more likely particular test results are in patients with the condition than in those without it. For example, if the likelihood ratio is 10 for a test result, the probability of a person having that result in the population with the condition is 10 times the probability of a person having the same result in the population without the condition. Likewise, if the likelihood ratio is 0.01 for a test result, the probability of having that result in the study population with the condition is 100th that of having the same result without the condition. Thus, the farther a likelihood ratio deviates from a value of one (equal probabilities in the affected and unaffected populations), the more informative the test result corresponding to that likelihood ratio becomes. Examples of the calculation of positive likelihood ratio values are given in Figure 2 using the hypothetical distributions shown.

Likelihood ratios < 0.1 or > 10 have been described as providing convincing diagnostic evidence, whereas those < 0.2 or > 5 give strong diagnostic evidence [29]. However, these guideline figures will not apply when pre-test suspicions of the presence of the condition are very high or very low. In these situations, likelihood ratios with values tending towards the low or the high extremes will be needed to rule out or rule in diagnoses, respectively.

With use of Bayes' theorem, likelihood ratios can be directly applied to derive probabilistic statements regarding the likelihood of a condition in an individual [28]. When applying this approach, likelihood ratios allow computation of *post-test* probabilities by the following formula:

Post-test odds = pre-test odds × likelihood ratio

Suppose that the rate of infertility in outpatients is 5%, and the likelihood ratio for a specific sperm concentration from an individual is 6 (see Figure 2), then

Pre-test odds = prevalence/(1 – prevalence)
= 0.05/0.95 = 0.053

Applying Bayes' theorem to the pooled negative likelihood ratio:

Post-test odds = pre-test odds × likelihood ratio
= 0.053 × 6 = 0.32

Converting post-test odds to *post-test* probability:

Post-test probability of infertility = post-test odds/(1 + post-test odds) = 0.32/(1 + 0.32) = 0.24 or 24 percent.

Such an approach would allow more complete use of the quantitative information available from semen analysis and would give an associated probability of having an outcome of infertility.

Difficulties commonly arise when trying to apply Bayes' theorem to update pre-test probabilities using the results of several tests, due to the non-independence of the test results. Albert [28] has described an approach in which the likelihood ratio for a combination of test results can be estimated using logistic regression. Although the contributions of individual tests to the final likelihood ratios cannot be discerned, the approach allows overall likelihood ratios for all possible combinations of test results to be deduced. The Albert model also allows inclusion of continuous test results dependent on assumptions being made about linearity. The derivation of the Albert model uses the data from a diagnostic accuracy study. Using as a 'training set' the test values from each person in a group of patients with and without the condition, logistic regression analysis can be applied to derive maximum likelihood estimates for the likelihood ratios. An example of this multivariate approach as applied to gestational monitoring is given in Boyd [30]. It is commonly recommended that the performance of models produced using these approaches be validated in new data sets.

6.3 More global multivariate models

For a more global prediction of fertility at the level of the couple, employing both information from the female partner and data from semen analysis, Hunault *et al*. [31] have developed two models of time to pregnancy based on Cox regression analysis of maternal factors, test results of semen analysis, referral status, and for one of the models, the results of post-coital testing. These models have been evaluated in an independent data set and seem to be useful in the evaluation of subfertile couples [32]. Such studies show the potential value of multivariate data analyses that incorporate the results of semen analysis with additional data.

7 Summary

This article provides a brief review of the various approaches that may be utilized for the analysis of human semen test results. Improvements in and standardization of the techniques for collection and analysis of human semen, as provided in the latest edition of the WHO manual should help reduce between-laboratory variability of semen analysis results [1]. The development of reference values for these results as provided by data from the large, multicentre reference interval study carried out by Cooper *et al*. [2] will help in efforts to standardize the interpretation of semen analysis. Although reference limits and decision limits derived for individual test results of semen analysis will undoubtedly be the tools of choice in the short term, in the long term, multivariate likelihood ratio methods for the interpretation of semen analysis alone or more complex predictive models that combine information from semen analysis with data derived from the female partner seem to represent better means for assessing the likelihood of achieving a successful pregnancy in a subfertile couple.

References

1 World Health Organization. WHO Laboratory Manual for the Examination and Processing of Human Semen, 5th ed. Geneva: World Health Organization; 2010.
2 Cooper TG, Noonan E, von Eckardstein S, Auger J, Baker HW, *et al*. World Health Organization reference values for human semen characteristics. Hum Reprod Update. 2009 Nov 24. [Epub ahead of print].
3 Solberg HE, Grasbeck R. Reference values. Adv Clin Chem 1989; 27: 1–79.
4 Young DS. Determination and validation of reference intervals. Arch Pathol Lab Med 1992; 116: 704–9.
5 Harris EK, Boyd JC. Statistical Bases of Reference Values in Laboratory Medicine. New York: Marcel Dekker, Inc.; 1995.
6 Horn PS, Pesce AJ. Reference Intervals: A User's Guide. Washington, DC: American Association for Clinical Chemistry; 2006.
7 Clinical and Laboratory Standards Institute (CLSI). Defining, Establishing, and Verifying Reference Intervals in the Clinical Laboratory; Approved Guideline-Third Edition CLSI document C28-A3 (ISBN 1-56238-663-8). Wayne: Clinical and Laboratory Standards Institute; 2008.
8 International Federation of Clinical Chemistry, Expert Panel on Theory of Reference Values. Approved recommendation on the theory of reference values: Part 1. The concept of reference values. J Clin Chem Clin Biochem 1987; 25: 337–42.
9 International Federation of Clinical Chemistry, Expert Panel on Theory of Reference Values. Approved recommendation on the theory of reference values: Part 2. Selection of individuals for the production of reference values. J Clin Chem Clin Biochem 1987; 25: 639–44.
10 International Federation of Clinical Chemistry, Expert Panel on Theory of Reference Values. Approved recommendation on the theory of reference values: Part 3. Preparation of individuals and collection of specimens for the production of reference values. J Clin Chem Clin Biochem 1988; 26: 593–8.
11 International Federation of Clinical Chemistry, Expert Panel on Theory of Reference Values. Approved recommendation on the theory of reference values: Part 4. Control of analytical variation in the production, transfer and application of reference values. Eur J Clin Chem Clin Biochem 1991; 29: 531–5.
12 International Federation of Clinical Chemistry, Expert Panel on Theory of Reference Values. Approved recommendation on the theory of reference values: Part 5. Statistical treatment of collected reference values: determination of reference limits. J Clin Chem Clin Biochem 1987; 25: 645–56.
13 International Federation of Clinical Chemistry, Expert Panel on Theory of Reference Values. Approved recommendation on the theory of reference values: Part 6. Presentation of observed values related to reference values. J Clin Chem Clin Biochem 1987; 25: 657–62.
14 Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults. Executive summary of the Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). JAMA 2001; 285: 2486–97.
15 Murphy EA. The normal. J Epidemiol 1973; 98: 403–11.
16 Langlois MR, Delanghe JR. Biological and clinical significance of haptoglobin polymorphism in humans. Clin Chem 1996; 42: 1589–600.
17 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.
18 Cinar C, Yazici C, Ergünsu S, Beyazyürek C, Javadova D, *et al*. Genetic diagnosis in infertile men with numerical and constitutional sperm abnormalities. Genet Test 2008; 12: 195–202.
19 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.
20 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.
21 Linnet K. Two stage transformation systems for normalization of reference distributions evaluated. Clin Chem 1987; 33: 381–6.
22 Solberg HE. RefVal: a program implementing the recommenda-

tions of the International Federation of Clinical Chemistry on the statistical treatment of reference values. Comput Methods Programs Biomed 1995; 48: 247–56 (programme is available from the author).
23 Horn PS, Pesce AJ, Copeland BE. A robust approach to reference interval estimation and evaluation. Clin Chem 1998; 44: 622–31.
24 Boyd JC. Reference regions in two or more dimensions. Clin Chem Lab Med 2004; 42: 739–46.
25 Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem. 1993; 39: 561–77.
26 Gunalp S, Onculoglu C, Gurgan T, Kruger TF, Lombard CJ. A study of semen parameters with emphasis on sperm morphology in a fertile population: an attempt to develop clinical thresholds. Hum Reprod 2001; 16: 110–4.
27 Nasr-Esfahani MH, Razavi S, Mardani M. Relation between different human sperm nuclear maturity tests and *in vitro* fertilization. J Assist Reprod Genet 2001; 18: 219–25.
28 Albert A. On the use and computation of likelihood ratios in clinical chemistry. Clin Chem 1982; 28: 1113–9.
29 Jaeschke R, Guyatt GH, Sackett DL. Evidence-Based Medicine Working Group. Users' guides to the medical literature. VI. How to use an article about a diagnostic test. B: What are the results and will they help me in caring for my patients? JAMA 1994; 271: 703–7.
30 Boyd JC. Mathematical tools for demonstrating the clinical usefulness of biochemical markers. Scand J Clin Lab Invest Suppl 1997; 227: 46–63.
31 Hunault CC, Habbema JD, Eijkemans MJ, Collins JA, Evers JL, *et al*. Two new prediction rules for spontaneous pregnancy leading to live birth among subfertile couples bases on the synthesis of three previous models. Hum Reprod 2004; 19: 2019–26.
32 Hunault CC, Laven JS, van Rooij IA, Eijkemans MJ, te Velde ER, *et al*. Prospective validation of two models predicting pregnancy leading to live birth among untreated subfertile couples. Hum Reprod 2005; 20: 1636–41.