Discrimination, adjusted correlation, and equivalence of imprecise tests: application to glucose tolerance

JONATHAN LEVY, RICHARD MORRIS, MARGARET HAMMERSLEY, AND ROBERT TURNER
Diabetes Research Laboratories, Department of Clinical Medicine, Oxford University, Oxford OX2 6HE, United Kingdom

Levy, Jonathan, Richard Morris, Margaret Hammersley, and Robert Turner. Discrimination, adjusted correlation, and equivalence of imprecise tests: application to glucose tolerance. Am. J. Physiol. 276 (Endocrinol. Metab. 39): E365–E375, 1999.—Comparison studies between physiological tests are often unsatisfactory for assessing their ability to distinguish between subjects. We recommend a simple but comprehensive protocol, using duplicate testing, that compares tests using 1) the discriminant ratio (DR) between the underlying between- and within-subject SDs, 2) correlation coefficients adjusted for attenuation due to test imprecision, and 3) unbiased estimation of the underlying linear relationship between test results. The following five alternative methods for assessing glucose tolerance were compared: fasting plasma glucose (FPG) as a single sample or as the mean of three 5-min samples (FPG3); the 1- and 2-h glucose during a low-dose intravenous glucose infusion (CIG); and the 2-h plasma glucose from a 75-g oral glucose tolerance test (OGTT). All tests had similar DRs ranging from 2.6 to 4.2. The adjusted correlation between FPG and CIG tests approached unity, and those between OGTT and other tests were ~0.9, showing that FPG provides similar information to the OGTT. FPG concentrations of 6.0 and 7.1 were found equivalent to the 1985 World Health Organization OGTT thresholds for impaired glucose tolerance and diabetes (7.8 and 11.1 mmol/l).

plasma glucose; precision

COMPLEX PHYSIOLOGICAL responses, such as glucose tolerance, may be assessed by many different methods. For instance, the control of plasma glucose may be measured by the fasting plasma glucose (FPG), by the response to an oral or an intravenous glucose challenge, by the percentage of glycated hemoglobin ($HbA_1c$), or by the plasma concentration of fructosamine. Different clinical tests are used to assess the same physiological variable because of different clinical circumstances, the application of advances in understanding and technique, or as a result of historical circumstances or fashion. The tests may assess differing or equivalent aspects of a physiological characteristic and may express their results in different units of measurement.

In this paper, we recommend a simple but comprehensive methodology for comparing different tests of a continuous physiological variable, which may be applied irrespective of the scales of measurements used. For such comparisons, it is important to include assessment of both the between- and within-subject variation of each test, as this allows comparison of the ability of tests to discriminate between different subjects, determination of the underlying correlation between tests after adjusting for attenuation due to within-subject variation, and unbiased estimation of the underlying relationships between the results of the different tests.

Omission of any of these components in a comparison study will provide inadequate information for choosing a particular test for a particular situation. The imprecision, or the within-subject variation, of a test is of little use by itself and must be considered in relation to the range of the test results. The “coefficient of variation,” which relates imprecision to the midpoint of the range, is often inappropriate as it does not take account of the dynamic range of a test (1) and is not always comparable between different scales of measurement. The present paper introduces the concept of “discrimination,” i.e., the ability to distinguish between individual subjects within a specified range of interest. This can be expressed as the discriminant ratio (DR), which is defined here as the ratio of the underlying between-subject SD ($SD_b$) to the within-subject SD ($SD_w$). The DR has a defined distribution, and DRs for different tests can be compared statistically.

The correlation between different tests must be included in any comparison. Test imprecision will diminish, or “attenuate,” the measured correlation coefficients in a well-described way (11), and it is important to correct for this so as to be able to assess the underlying “true” correlation, as this represents the degree to which the tests are assessing the same physiological trait. This attenuation adjustment can be expressed in terms of the DRs of the respective tests.

Finally, the comparison of tests must, if possible, relate measurements by one test to those by another. Where the relationship is linear, the gradient of the relationship obtained by least squares regression is underestimated (“regression dilution”) when both tests are subject to appreciable measurement error. An unbiased technique for deriving the relationship is therefore necessary, and, while many methods are available, a suitable method (27) is recommended here.

The assessment of glucose tolerance is a specific area where several methods are available to an investigator, for instance, by the measurement of steady-state plasma glucose or the response to a glucose challenge, either oral or intravenous. The standard test of glucose tolerance has, until recently, been the oral glucose tolerance test (OGTT; see Ref. 33), but, in practice, this test is not often performed. This is partly because of the inconvenience of a 2-h test and partly because of the marked variability of the OGTT. The poor reproducibility of the test (12, 21, 24, 26), which has a reported coefficient of variation of the order of 15–40%, due in part to the variable rate of gastric emptying, has predictable effects on recategorization of subjects on repeat tests, with a change in status on repeat testing in 30–60% of cases.
of impaired glucose tolerance (IGT; see Refs. 9, 13, 26, 28, 29) and predictable regression to the mean (26). The simple measurement of FPG has been suggested as a preferable measure (4, 8, 15, 20), and a continuous intravenous infusion of glucose (CIG) can assess glucose tolerance and give simultaneous measures of pancreatic β-cell function and insulin resistance (16). The FPG thresholds for diabetes have been set using outcome data (4). We evaluate and compare the performance of this and other tests throughout the physiological range.

This paper outlines the concepts and components of a physiological comparison study and uses as an example the assessment and comparison of different tests for measuring glucose tolerance by either 1) the FPG (single sample or mean of 3 consecutive samples), 2) the 1- and 2-h responses to a CIG, and 3) the standard 2-h response to the OGTT, in repeated tests in 30 subjects spanning the range of glucose tolerance.

METHODS

Statistical Methods

This paper considers the following three aspects relating to the assessment and comparison of different tests for measuring an underlying physiological variable such as glucose tolerance: 1) the ability of a test to discriminate between different subjects and comparison of discrimination between different tests; 2) the correlation between pairs of tests, adjusting for bias due to within-subject variation. Such variation attenuates measured correlation coefficients so that they underestimate the underlying true correlation. This is important in assessing the degree to which different tests purporting to assess the same underlying trait differ with respect to systematic between-subject factors as opposed to random within-subject variation; and 3) in cases where the relationship between a pair of tests appears to be a linear, unbiased estimation of the DR is calculated simply as

\[ DR = \sqrt{(\text{SD}_B^2 - \text{SD}_W^2/k)}/\text{SD}_W \]

This result may also be obtained by an analysis of variance (ANOVA) approach, using a fixed effects model, and this is presented in Appendix I. The appropriate equation is then

\[ DR = \sqrt{\left[\frac{(k \times \text{MS}_B - \text{MS}_W)}{k} \right]} \]

where \( \text{MS}_B \) is the between-subject mean square, \( \text{MS}_W \) is the within-subject mean square, and \( k \) is the number of replicate tests in each subject, as above.

Assuming that the within-subject variation is normally distributed, we have used an analytical approach to derive confidence limits for estimated DR values and to test for the significance of differences between the DRs of different tests. We present these in Appendix I, where we also discuss our methodology in relation to alternative measures of test “reliability,” in particular the intraclass correlation coefficient (ICC; see Refs. 5, 30).

Correlation between pairs of tests. Two tests designed to assess a complex physiological characteristic, such as degree of glycemic control, may use different methodologies, neither of which may perfectly represent the characteristic in question. The results of the tests may differ in systematic ways, independently from their random within-subject variation. The degree to which the tests measure the same characteristic may be assessed by the correlation between their results in a set of subjects. In the absence of within-subject variation, the degree to which their correlation falls short of unity would represent the extent to which the tests either fail to measure precisely the same characteristic or to measure aspects of the underlying characteristic that are differentially influenced by other factors that vary between subjects. It is this underlying true correlation that we are interested in here. Test imprecision, however, further attenuates the observed correlation so that imperfect correlation will usually be due to a combination of the systematic between-subject factors discussed above and the presence of within-subject variation.
A study comparing tests must distinguish between these two components, and this can be achieved using a standard formula that corrects measured correlations for attenuation (11). Such a correction requires knowledge of both the within- and between-subject variation of the measured values. Because the DR incorporates both of these elements, the correction can be expressed in terms of the DR

\[
\eta = k \times \sqrt{\frac{\text{DR}^2}{1 + k \times \text{DR}^2}} \times \sqrt{\frac{\text{DR}^2}{1 + k \times \text{DR}^2}}
\]

(see APPENDIX I). The corrected correlation coefficient, therefore, represents the degree to which the tests represent the same physiological characteristic, independent of test imprecision.

Unbiased estimation of a linear relationship. Where two tests represent the same characteristic, their results will often be found to be linearly related (although this may require numerical transformation). The determination of this relationship will be important in relating the results of one test to those of the other. Within-subject variability will lead to a noisy relationship, and a statistical approach is necessary to estimate the underlying linear equation. Although least-squares linear regression is often used for this, it is not appropriate here since it assumes that the explanatory variable is free from noise. When this is not the case, linear regression will underestimate the slope of the equation, a well-recognized effect termed regression dilution.

There is no perfect method of estimating the true relationship in these circumstances, but the method chosen here is that of least perpendicular distances corrected for scale differences, which has been shown to perform well in relation to other methods (27). The equations for this are presented in APPENDIX I.

Design of comparison studies. There is no single measure that can be used to compare physiological tests. Assessment and comparison of test discrimination and the determination of their underlying correlations and inter-relationships are equally important components. These all require simultaneous consideration of both between-subject and within-subject variation within the physiological range of interest. A comparison study should, ideally, assess both of these factors by using replicate tests in subjects chosen to span that range.

Experimental Protocols

Subjects. Thirty white Caucasian subjects were studied, consisting of 10 normoglycemic subjects, 9 subjects with IGT, and 11 with type II diabetes according to 1985 World Health Organization (WHO) definitions (33). All subjects were on a weight-maintaining diet and had not changed their medical therapy for 4 wk before the tests. Subject characteristics are presented in Table 1 by glucose tolerance test group.

Protocols. Each subject was studied on four occasions within a 6-wk period. After a 12-h fast, subjects went to the hospital and sat on a bed for the duration of the tests. Tests were each performed on two occasions in the same subject and in random order.

FPG and CIG. Two cannulas were inserted in the same arm. One, for blood sampling, was placed at the wrist or on the dorsum of the hand, which was heated by an electrical blanket to “arterialize” the venous blood. The other cannula, for infusion of glucose, was placed in an antecubital vein. A blood sample was taken at time 0, and the plasma glucose concentration at this single time point was termed FPG1. Blood samples were also taken at times 5 and 0 min, and the mean of the plasma glucose at the three time points was termed FPG2. At time 0, a continuous 5 mg·kg⁻¹·min⁻¹ infusion (22) of 10% glucose was started and continued for 2 h. One-hour CIG and 2-h CIG glucose were defined as the means of the three plasma glucose concentrations in blood samples taken at 50, 55, and 60 min and at 110, 115, 120 min, respectively.

OGTT. A single cannula was placed in an antecubital vein for blood sampling. Fasting blood samples were taken at 0, 5, and 0 min. At 0 min, the subject consumed a 75-g glucose drink, and blood samples were taken at 30, 60, 90, and 120 min.

Biochemical Assays

Plasma glucose was determined by a hexokinase-based method (Boehringer Mannheim UK, Lewes, UK) on a centrifugal COBAS MIRA autoanalyzer (Roche, Welwyn Garden City, UK).

RESULTS

Estimates of Glucose Tolerance

Values for glucose tolerance using FPG1, FPG3, 1-h CIG, 2-h CIG, and 2-h OGTT are presented in Table 2 as median and ranges. Fasting and CIG measures were

Table 1. Subject characteristics

<table>
<thead>
<tr>
<th></th>
<th>Normal</th>
<th>IGT</th>
<th>Diabetes</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>10</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>Sex (M/F)</td>
<td>5/5</td>
<td>4/5</td>
<td>8/3</td>
</tr>
<tr>
<td>Age, yr (range)</td>
<td>41 (19–64)</td>
<td>50 (25–63)</td>
<td>58 (24–67)</td>
</tr>
<tr>
<td>Body mass index, kg/m²</td>
<td>26.0 ± 4.7</td>
<td>28.3 ± 9.6</td>
<td>28.5 ± 3.8</td>
</tr>
<tr>
<td>Fasting plasma glucose, mg/dl (mmol/l)</td>
<td>86 ± 9</td>
<td>112 ± 14</td>
<td>140 ± 13</td>
</tr>
<tr>
<td>2-h OGTT, mg/dl (mmol/l)</td>
<td>106 ± 23</td>
<td>142 ± 40</td>
<td>259 ± 38</td>
</tr>
<tr>
<td>(5.9 ± 1.3)</td>
<td>(7.9 ± 2.2)</td>
<td>(14.4 ± 2.1)</td>
<td></td>
</tr>
</tbody>
</table>

Values are means ± SD; n, no. of subjects. IGT, impaired glucose tolerance; M, male; F, female; OGTT, oral glucose tolerance test.

Table 2. Within- and between-subject SDs and discriminant ratios of measures of glucose tolerance

<table>
<thead>
<tr>
<th></th>
<th>Median (Range)</th>
<th>Corrected Between-Subject SD</th>
<th>Within-Subject SD</th>
<th>Discriminant Ratio (1 SE range)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FPG1, mmol/l</td>
<td>6.2 (3.8–11.5)</td>
<td>1.37</td>
<td>0.53</td>
<td>2.6 (2.2–3.0)</td>
</tr>
<tr>
<td>FPG3, mmol/l</td>
<td>6.4 (3.8–11.6)</td>
<td>1.39</td>
<td>0.42</td>
<td>3.3 (2.9–3.9)</td>
</tr>
<tr>
<td>1-h CIG, mmol/l</td>
<td>10.4 (6.0–14.4)</td>
<td>2.15</td>
<td>0.71</td>
<td>3.0 (2.6–3.5)</td>
</tr>
<tr>
<td>2-h CIG, mmol/l</td>
<td>11.3 (5.6–15.5)</td>
<td>2.87</td>
<td>0.68</td>
<td>4.2 (3.6–4.8)</td>
</tr>
<tr>
<td>2-h OGTT, mmol/l</td>
<td>9.4 (3.5–18.4)</td>
<td>0.95 (0.54–1.26)</td>
<td>0.19</td>
<td>3.2 (2.7–3.7)</td>
</tr>
</tbody>
</table>

FPG1, single sample fasting plasma glucose; FPG3, fasting plasma glucose from 3 samples; CIG, continuous iv glucose infusion.
homoscedastic, and the within- and between-subject variations are illustrated as plots of difference (first test—second test) vs. mean (of the 2 tests) in Fig. 1. The 2-h OGTT was found to have within-subject variation increasing with mean values, and this was corrected by log transformation, with the transformed data presented as difference vs. mean plots in Fig. 2 and as medians and ranges for the whole group in Table 2. The underlying SD_U and SD_W and the DRs of the tests are also presented in Table 2.

DR values for the five measures, together with the one SE range of the estimates, are illustrated in Fig. 3. Although the lowest DR was the FPG_1 and the highest the 2-h CIG, there was no significant difference between them on the overall statistical test ($\chi^2 = 6.2, P = 0.19$, using Eq. 10 in APPENDIX I). Consideration was given to the exclusion of a subject whose inter-test difference in the 2-h CIG was four SDs from the mean of the rest of the group (see Fig. 1). In the absence of an identifiable reason for the large difference between his two test values, this subject was included in the analyses presented here, although, if he were excluded, the DR of the 2-h CIG would rise to 6.1, significantly greater than the DRs of the other tests.

Table 3 shows the Pearson correlations between the tests, both before and after adjustment for attenuation. The correlations were calculated between subject means of duplicate tests, to give the best estimate of the underlying relationships between tests. Adjusted correlations between the fasting and both of the intravenous measures were high, approaching one. Those between the 2-h OGTT results and the other tests were somewhat lower (~0.9), indicating that there was some biological discordance in their relationship, independent of their within-subject measurement error.

Figure 4 shows the scattergram between the 2-h plasma OGTT (on a logarithmic scale) and the FPG. It also illustrates the line of equivalence, derived as explained above, and the linear regression line of the 2-h OGTT on FPG. The dilution effect of the within-subject variation on the regression line can be seen clearly. Table 4 gives coefficients for the unbiased linear equations relating the test values, and Table 5 gives the points on the various scales that are equivalent to the 1985 WHO OGTT thresholds for IGT and diabetes.

SD_W of the logarithm of the 2-h OGTT may be interpreted in relation to a standard interval in the range of glucose tolerance, for instance, the interval between the thresholds for IGT (7.8 mmol/l) and for diabetes (11.1 mmol/l). This interval is 0.153 on a logarithmic scale (base 10), whereas the SD_W is 0.060. The difference between two individual measurements at either end of this interval would not be significant at the 5% level ($P = 0.07$), given the imprecision of the 2-h OGTT. In other words, it would not be possible to confidently distinguish between two such individual values. The same applies to all other tests examined, using the unbiased equivalent values to these thresholds given in Table 5, apart from the 2-h CIG plasma glucose, for which the two measurements at opposite ends of the equivalent interval would differ significantly ($P = 0.013$).

**DISCUSSION**

This study has shown that different methods of assessing glucose tolerance were broadly comparable in a range of subjects spanning normal glucose tolerance,
IGT, and type II diabetes. This assessment involved the following three separate components: 1) comparison of the discrimination of the tests, i.e., their ability to distinguish between different subjects, 2) determination of the degree to which different tests measure the same underlying physiological property, and 3) estimation of the underlying relationship between the test results. The assessment of the within-subject imprecision of each test is a fundamental requirement for this evaluation, so that comparison studies must involve at least duplicate measurements in all subjects in order to determine both the between- and within-subject measurement variation of each test in the same group of individuals. This is best done in subjects who represent a clinically meaningful range of glucose tolerance. To ascribe a single value to within-subject imprecision requires homoscedasticity, and numerical transformation of results may be necessary to achieve this, as illustrated by the 2-h plasma glucose from the OGTT. A measure of imprecision is important for the assessment of changes within subjects, such as over time or after interventions (10).

The determination of imprecision alone is not adequate, however, for the assessment of the practical value of a test. The standard methods of assessing imprecision, including the coefficient of variation, have little meaning on their own, without reference to the range of measurements to which they are being applied. The ability to distinguish between individuals within this range is here termed the discrimination of the test and is assessed using the DR.

The concept of discrimination should be distinguished from the ability of a test to categorize patients by an external gold-standard dichotomy. This is a particular concern in the field of clinical chemistry, for instance, when a biochemical test is being assessed for its ability to detect the presence of a malignancy. Receiver operating characteristics (34) have been used for this purpose. However, they are not suitable for the assessment of test results as a continuous scale of measurement or for the comparison of tests without reference to an external categorization. In the field of glucose tolerance, categories have been defined on the basis of thresholds in a continuous scale of measurement for the OGTT based on external criteria, but this is a notoriously imprecise test (7), and it would not therefore be appropriate to assess possible alternative tests using a categorical approach based on these thresholds. The use of the DR provides a means of comparing how well the subjects studied can be reliably distinguished by different tests, which is an important component of a comprehensive comparison of imprecise tests. For instance, in many research studies using continuous variables, the statistical power to distinguish between groups of subjects or to determine correlations between variables will depend on discrimination.

A very similar concept, referred to as reliability, has been assessed, particularly in the psychological literature, using the ICC. This relates the covariance of

---

Table 3. Measured Pearson correlation coefficients between within-subject means of duplicate tests

<table>
<thead>
<tr>
<th></th>
<th>FPG</th>
<th>1-h CIG</th>
<th>2-h CIG</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-h CIG</td>
<td>0.93 (0.98)</td>
<td>0.91 (0.94)</td>
<td>0.97 (1.0)</td>
</tr>
<tr>
<td>2-h CIG</td>
<td>0.85 (0.89)</td>
<td>0.83 (0.88)</td>
<td>0.87 (0.90)</td>
</tr>
</tbody>
</table>

Values adjusted for attenuation are in parentheses.
replicate test results to the combined between-subject and within-subject variance and is algebraically related to the DR. However, in the context of discriminating between different subjects, it is not as easy to grasp as the DR, which has a direct intuitive relevance when considering a test's application. Furthermore, it is not as easy to derive statistical tests comparing different ICC values. A recently described method for the comparison of two ICCs does not extend to the comparison of more than two tests and was only validated for studies employing repeated tests in 100 subjects or more (2).

The choice of a methodology for comparing different tests is intimately related to its theoretical basis and in particular to the availability of statistical criteria for assessing such a comparison. Much of the theoretical discussion of the ICC and treatments of comparative tests between ICCs have been based on random effects models in which the individuals in which the tests are being performed are assumed to be drawn at random from a known normally distributed population. This presupposes a particular population in which the tests are being applied. Complex populations (for instance those consisting of subgroups) would require appropriately complex statistical models. Unfortunately, even the simplest random effects models are hard to treat analytically, as in the case of the ICC quoted above. Our approach has been to concentrate on evaluating tests across a particular physiological spectrum of interest. In the example presented here, for instance, glucose tolerance represents a physiological and pathological

Table 4. Gradient and intercept coefficients for lines of equivalence

<table>
<thead>
<tr>
<th>y-Variable</th>
<th>FPG₃</th>
<th>1-h CIG</th>
<th>2-h CIG</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-h CIG, mmol/l</td>
<td>1.54, 0.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2-h CIG, mmol/l</td>
<td>2.09, −2.32</td>
<td>1.33, −2.59</td>
<td></td>
</tr>
<tr>
<td>Log 2-h OGTT</td>
<td>0.14, 0.071</td>
<td>0.090, 0.024</td>
<td>0.065, 0.23</td>
</tr>
</tbody>
</table>

Line of equivalence determined as follows: \( y = 1.54x + 0.40 \).

Table 5. Equivalent values to common OGTT thresholds

<table>
<thead>
<tr>
<th>OGTT Thresholds, mmol/l</th>
<th>Fasting plasma glucose</th>
<th>1-h CIG plasma glucose</th>
<th>2-h CIG plasma glucose</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.8</td>
<td>6.0</td>
<td>9.6</td>
<td>10.1</td>
</tr>
<tr>
<td>11.1</td>
<td>7.1</td>
<td>11.3</td>
<td>12.5</td>
</tr>
</tbody>
</table>

unity irrespective of the its distribution in particular populations. For this purpose, we take the view that it is appropriate to perform a comparison study in subjects selected to span the range of interest, which may be analyzed by a fixed-effects statistical model. Such an analysis is presented here for the DR, allowing the derivation of straightforward expressions for the SD and confidence intervals of the DR and the evaluation of the statistical significance of differences between the DRs of different tests.

Although the comparison of the DRs of different tests in the same study is valid, the DR in itself is not a universal characteristic of a test, as it will depend on the choice of subjects on which the comparison is performed. When the subjects cover a greater range of values, the DR will be larger, and vice versa, and the DR calculated when subjects are selected to span a range of interest will generally be larger than if subjects had been chosen randomly from the same population. However, it is a fundamental property when it comes to a test's practical application, and, unlike imprecision, it can be used as a basis for comparison between tests. The DRs of the five tests examined here were not significantly different, in spite of the increased complexity and expense of the CIG and the OGTT.

Two perfectly precise tests assessing the same physiological variable would be perfectly correlated. Departures from perfect correlation can be due to the following two factors: 1) underlying differences not directly related to the variable of interest. These will manifest themselves as systematic differences between subjects; for instance, in the assessment of glycemic control, which is determined by the FPG and OGTT in qualitatively different ways, the underlying correlation may fall short of unity because of the influence of factors such as the effect of gastric emptying and the influence of intestinal incretin effects that differ between the two methodological approaches; 2) diminution of the underlying correlation may also arise from within-subject variation, and this is a well-described statistical effect termed “attenuation,” which may be adjusted for by standard techniques to enable the estimation of the DR and the OGTT.

Adjustment for attenuation will establish the degree to which the underlying correlation differs from unity due to the factors detailed in factor 1 above.

In this paper, the adjusted correlation coefficients between the fasting glucose and the CIG approached
unity and were only slightly lower between these tests and the 2-h OGTT. Although additional factors unrelated to the homeostatic control of plasma glucose, such as variable gastric emptying, would contribute to this, the relatively high overall intercorrelations and the simplicity and cheapness of the FPG would recommend this as the measure of choice for the assessment of glucose tolerance.

When two tests have an underlying structural relationship between their measurements (or transformations of these) that is linear, it can be instructive to determine the equation of the "line of equivalence." Linear regression, although often used, is unsatisfactory since it assumes perfect precision in the independent variable and is subject to regression dilution. There have been many approaches to deriving an unbiased estimation, as comprehensively reviewed by Riggs et al. (27), and the "weighted least squared perpendicular distance" approach (Riggs' "PW" method) has been used in this paper. In the present study, the assumption of linearity, within the limits set by the imprecision of the tests, was supported by visual inspection of plots of the relationships (data not shown). The FPG threshold concentrations recommended by the American Diabetes Association (4), based on studies of the prevalence of retinopathy in three distinct populations, were confirmed as equivalent to the established OGTT thresholds for IGT and diabetes.

We also calculated the SD of log_{ed}(DR) estimates for different numbers of replicate tests, using the Taylor series expansion, subject to a constraint on the total number of tests performed. For a study comparing two methods using a total of 60 tests, the power to detect a difference between DRs of 2.5 and 4.0 is 58% using 30 subjects and two replicate tests, rising to 72% with 20 subjects and three tests, an increase in power of 24%. Further increasing the number of replicates gives smaller increases in power, e.g., for four tests in 15 subjects the power is 78%, with even smaller gains for more than four replicates. There would appear to be some advantage in using three, rather than two, replicate tests in each subject, but there is little advantage in increasing the number beyond three given the need to obtain sufficient subjects to adequately cover the range of interest.

This study showed that, with OGTT done under carefully controlled conditions, with a reproducibility that is somewhat better than reported elsewhere, it was not possible to distinguish between the WHO thresholds for IGT and diabetes at a 5% significance level, and this was also the case for FPG, even when the mean of three samples at 5-min intervals was assessed (the between-sample variation being relatively small in relation to the between-day variation). It is therefore not surprising, with two thresholds close together, that repeat measurements often give change of status. Improved classification could be achieved by taking the mean of determinations on more than one day. However, although these classifications may be useful for epidemiological purposes, for practical purposes the actual OGTT or FPG value is more informative (31, 32).

In summary, we have outlined a comprehensive but simple methodology for the comparison of imprecise tests, encouraging 1) comparison of test discrimination, expressed as the DR, 2) the evaluation of the degree of agreement between tests based on correlation coefficients adjusted for attenuation, and 3) in the case of a linear relationship between test results (or their mathematical transformations), the use of an unbiased method for estimating the underlying equation. For such a comparison study, it is important to determine the within-subject variation of each test as well as the variation between subjects. Application of these methods to various tests of glucose tolerance demonstrated similar discrimination, acceptable agreement, and an unbiased estimation of the FPG values equivalent to those of the 2-h OGTT. The latter agree closely with the outcome-derived thresholds currently being recommended by the American Diabetes Association. However, because the thresholds for IGT and diabetes are within measurement error and cannot be reliably distinguished, the absolute 2-h OGTT or FPG is more informative than the categorization.

APPENDIX I: STATISTICAL METHODS

This section presents a more detailed mathematical treatment of the concepts outlined in METHODS.

Discrimination Between Subjects

Statistical model. We consider the comparison of different tests, each measuring the same physiological variable. Each test is performed k times on each of n subjects, with the order of the tests being randomized for each subject.

Considering first a single test in isolation, an appropriate model is

\[ X_{ij} = \mu + \alpha_i + \epsilon_{ij} \quad \text{for } i = 1, \ldots, n \text{ and } j = 1, \ldots, k \]  

where \( X_{ij} \) is the result of the test performed for the \( j \)th time on the \( i \)th subject, \( \mu \) is the overall mean value of the variable in question on the scale of the current test, and \( \alpha_i \) is the true value of the \( i \)th subject, measured as a deviation from the mean (thus \( \sum_{i=1}^{n} \alpha_i = 0 \)). \( \epsilon_{ij} \) represents day-to-day variation, which includes both biological and assay variation; the \( \epsilon_{ij} \) are assumed to be independent, normally distributed random variables with mean zero and variance \( \sigma^2 \). Equation 1 is a standard one-way ANOVA.

The assumption of constant variance (or homoscedasticity) of the error term, \( \sigma^2 \), can be checked graphically. If \( k > 5 \), we can calculate the quartiles of the \( k \) replicate test results for each subject and plot \( \log(\text{interquartile range}) \) against \( \log(\text{median}) \) (14). If \( 2 < k < 5 \), plot the SD of the \( k \) replicates against the mean for each subject (23), and if \( k = 2 \) plot the differences (1st - 2nd replicate) between the pairs of tests against the subject means (6). If the assumption of homoscedasticity holds, the plotted measure of variation [log(interquartile range), SD, or difference] should be approximately constant across the range of subjects. If there appears to be a systematic relationship between the measure of variation and subject medians or means, this can often be removed by mathematically transforming the results of \( X_{ij} \). A common case in physiological measurements is where the SD increases in direct proportion to the mean, when a log transformation of the \( X_{ij} \) stabilizes the variance and \( \log(X_{ij}) \) can then be used in place of \( X_{ij} \) in Eq. 1. Other transformations can be
considered for different relationships between the subject SDs and means (14, 23).

It is also possible to check the assumption that the $e_{ijh}$ have a normal distribution by plotting the ordered residuals from fitting Eq. 1 against standard normal deviates in a "normal probability plot" (3). However, the ANOVA procedures used here are fairly robust to moderate departures from normal distribution and can be used without such sophisticated checking, provided homoscedasticity of variance holds and the data do not exhibit marked skewness.

In these experiments, subjects are selected to span a range of glucose tolerance and are not chosen randomly from a prespecified population. The subject effects $\alpha_i$ are therefore considered as "fixed" rather than "random" effects.

**DR.** As a measure of discrimination between subjects, we define the true DR, $\Delta$, as the ratio of the underlying SD$_B$ to the SD$_W$

$$\Delta = \sqrt{\frac{\sum_{i=1}^{n} \alpha_i^2 (n - 1)}{\sigma}} \quad (2)$$

Unbiased estimates of the between- and within-subject variances are given by $(MS_B - MS_W)/k$ and $MS_W$, respectively, where $MS_B$ and $MS_W$ are the between- and within-subject mean squares from a standard one-way ANOVA, i.e.

$$MS_B = \frac{k}{(n-1)} \sum_{i=1}^{n} (M_i - M)^2$$

$$MS_W = \frac{1}{(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{k} (X_{ij} - M_i)^2$$

and $M_i = \frac{1}{k} \sum_{j=1}^{k} X_{ij}$ and $M$ is the overall mean.

We then estimate $\Delta$ empirically as the ratio of the between- to within-subject standard deviations

$$DR = \frac{1}{\sqrt{\frac{(MS_B - MS_W)/(k \times MS_W)}}} \quad (3)$$

The DR is algebraically related to the ICC, which is commonly used as a measure of the reliability of tests (5, 30)

$$DR = \sqrt{ICC/(1 - ICC)} \quad (4)$$

However, the methodology developed for ICCs is in the context of a random effects model, rather than the fixed effects model used here, so published results for SDs and confidence intervals cannot be used. The DR gives a measure that is intuitively closer to the idea of discrimination between subjects, whereas the ICC is a measure of correlation. In addition, for tests with good discrimination, ICC values tend to cluster unhelpfully close to their upper limit of one. Furthermore, there is no simple practicable test available for the comparison of ICCs from different tests in a random effects model. A recently described method for the comparison of two ICCs does not extend to the comparison of more than two tests and was only validated for studies employing repeated tests in 100 subjects or more (2). We have derived straightforward expressions for the SD and confidence intervals of the DR in a fixed effects model and a test for the equivalence of DRs in a comparison study.

Confidence limits for DR. Confidence limits for the DR can be found by noting that

$$DR = \sqrt{\frac{(F_0 - 1)/k}{k}} \quad (5)$$

where $F_0 = MS_B/MS_W$ is the standard F statistic from the one-way ANOVA. $F_0$ has a noncentral F distribution with degrees of freedom $v_1 = n - 1$ and $v_2 = n \times (k - 1)$ and noncentrality parameter

$$\lambda = k \times \sum_{i=1}^{n} \alpha_i^2/\sigma^2 = (n - 1) \times k \times \Delta^2$$

$\lambda$ can be estimated by $(n - 1) \times k \times DR^2$, and a 95% confidence interval for $\Delta$ is

$$\sqrt{((F_L - 1)/k)} \leq \Delta \leq \sqrt{((F_U - 1)/k)} \quad (6)$$

where $F_L$ and $F_U$ are the lower and upper 2.5% of the noncentral F ($\lambda$, $v_1$, $v_2$).

Noncentral F tables are not widely available (18), and a reliable approximation to $F_L$ and $F_U$ can be made using a central F distribution (25)

$$F_L = (1 + k \times DR^2) \times F^{*} \quad and \quad F_U = (1 + k \times DR^2) \times F^{*}$$

where $F^{*}$ is the lower and upper 2.5% of a central $F$ distribution.

Comparison of DRs. We derived a test for the equivalence of several DRs by assuming the following model

$$X_{ijh} = \mu_h + \alpha_{ih} + e_{ijh}$$

for $i = 1, \ldots, n$, $j = 1, \ldots, k_r$ and $h = 1, \ldots, r$

$X_{ijh}$ is now the result of the $h$th test performed for the $j$th time on the $i$th subject, $\mu_h$ is the mean value of the variable in question on the scale of test $h$, and $\alpha_{ih}$ is the true value of the $i$th subject measured using test $h$ ($\Sigma_{i=1}^{n} \alpha_{ih} = 0$ for each test $h$); $e_{ijh}$ are again assumed to be independent, normally distributed random variables with zero mean and variance $\sigma^2_{e_{ijh}}$.

Using this model, the DRs for each test are statistically independent. We used simulations (see APPENDIX II) to show that the distribution of $\log(DR)$ is approximately normal if the model assumptions hold. We then used Cochran's theorem (19) to show that the statistic Q has a $\chi^2$ distribution with $r - 1$ degrees of freedom, where

$$Q = \sum_{h=1}^{r} w_h \times (L_h - L)^2$$

$$L_h = \log_{e}(DR \text{ of test } h) \quad and \quad L = \left|\sum_{h=1}^{r} w_h \times L_h\right| \left/ \sum_{h=1}^{r} w_h\right| \quad (7)$$

$w_h = 1/s^2_h$ and $s_h$ is the standard deviation of $L_h$.

The DRs are unequal at a significance level of 0.05 if $Q$ exceeds 95% of the $\chi^2$ distribution.

We derived an expression for $s_h$, the estimated SD of $L_h$, from the mean and variance of the noncentral F distribution using a Taylor series expansion; details are given in APPENDIX III.

Alternative models. The models we have used, given by Eqs. 1 and 9, for observations from the comparison study have been deliberately chosen for their relative simplicity. Although some of the algebra is intricate, all of the calculations we have presented can be easily implemented using spreadsheet software and do not require the use of specialized statistical packages. However, some of our models assumptions do warrant further discussion.
First, the choice of a fixed rather than a random effects model is unusual in this kind of context. However, subject selection in our study is clearly nonrandom in that we have deliberately chosen roughly equal numbers of normal glucose tolerance, IGT, and diabetic subjects. Even within each of these subpopulations, sampling is unlikely to be random as subjects are sought to span the range of interest as evenly as possible, which is likely to result in oversampling from the extremes of the distribution. Such a sampling scheme is likely to produce a DR that is higher than that which would be obtained from a random sample from the population, and its use is restricted to comparison with other tests in the same study. It is not appropriate to formally compare test DRs that have been derived from different populations.

A population consisting of clearly defined subpopulations might be best treated using a mixed model, or even structural equation modelling. This would require more sophisticated analytical techniques and a larger scale of comparison study than we have presented in this paper. In the particular example presented here, however, the “subgroups” are not clearly separable but are arbitrarily defined by thresholds in a continuous spectrum. In this situation, which is relatively common in physiology, the approach taken here would be adequate, relatively simple, and practical. A formal comparison of the use of more complex models and the simple approach made here is beyond the scope of this paper.

The assumption of independence of the error terms \( \epsilon_{ijh} \) is unlikely to be completely true since most biological measurements exhibit some degree of “autocorrelation,” i.e., correlation between successive measurements made on the same subject. In the context of these studies, where repeat measurements are almost always made on different days and often several days or even weeks apart, the magnitude of such autocorrelation is likely to be small compared with the total within-subject variation in which we are interested. Furthermore, accurately estimating autocorrelation coefficients would be difficult in relatively small studies, and the degree of mathematical complexity would increase such that specialized statistical methodology and software would be needed, rendering the procedures inaccessible to many researchers. However, our methodology might not be appropriate where repeat measurements were made within the same day or where other biological reasons existed for suspecting nonnegligible autocorrelation.

Correlation Between Pairs of Tests

The nature of the relationship between a pair of tests can be examined graphically by plotting the subject means for the first test against those for the second. In many cases, particularly after transformations to ensure homoscedasticity, the relationship will be approximately linear, and the degree of correlation can be assessed using the Pearson product-moment correlation coefficient, \( r \) (3).

In the model of Eq. 9 for \( r = 2 \) tests, we are interested in the correlation between the underlying subject means \( \alpha_1 \) and \( \alpha_2 \). However, in the presence of within-subject variation, the sample correlation coefficient, i.e., the correlation between the two sets of observed subject means, underestimates the true correlation between the tests; this effect is known as attenuation and means that, even if the true subject means \( \alpha_1 \) and \( \alpha_2 \) were perfectly correlated, the correlation between the observed subject means would be less than unity because of the random fluctuations due to within-subject measurement variation.

Standard results from measurement error theory (11) show that the correlation between two measurements, both of which are subject to error, is attenuated by the factor

\[
\eta = \sqrt{(k_2 \times k_2)}
\]

where \( k_1 \) and \( k_2 \) are the reliability coefficients of the two tests. From Eq. 6

\[
k = DRM^2/(1 + DRM^2)
\]

where DRM is the DR of the means \( M_i \), \( i = 1, ..., n \), of the \( k \) replicate measurements on each subject, rather than of the individual measurements themselves.

Taking the mean of Eq. 1 over the \( k \) replicates yields

\[
M_i = \mu + \alpha_i + \xi_i \quad \text{for} \ i = 1, \ldots, n
\]

where \( \xi_i \) is normally distributed with mean of zero and variance \( \sigma^2/k \). Thus, in Eq. 2, \( \sigma \) must be replaced by \( \sigma/\sqrt{k} \), which we estimate by \( \sigma/\sqrt{(MS_w/k)} \), and from Eq. 5

\[
DRM = DR \times \sqrt{k}
\]

Thus

\[
\eta = \sqrt{[DRM^2/(1 + DRM^2)] \times [DRM^2/(1 + DRM^2)]}
\]

The Pearson correlation coefficient \( r \) can be adjusted for attenuation by dividing it by \( \eta \), i.e.

\[
r_{adj} = r/\eta
\]

where \( r_{adj} \) is the adjusted \( r \).

In cases where the relationship between the tests is clearly nonlinear, the Spearman rank correlation coefficient \( r_s \) should be used in place of \( r \) to assess the comparability of the tests. However, there is no universal formula for the attenuation of \( r_s \) in the presence of measurement error.

Unbiased Estimation of Linear Relationship

In the case where the relationship between a pair of tests is linear, it may be useful to obtain unbiased estimates of the gradient and intercept of the line. Linear regression gives biased estimates because it only considers errors in the dependent variable, and clearly both tests here are subject to error; the gradient is always underestimated, and regression of subject means from test 1 on those from test 2 clearly gives a different relationship to that of test 2 on test 1.

The method that we have chosen to estimate the linear relationship between the subject mean measurements from test 1 and those from test 2 is that of “perpendicular least squares, properly weighted.” This essentially minimizes the sum of the squared perpendicular distances between the observed data and the fitted line, but with an adjustment that makes the method invariant to linear transformations of the measurement scales. If \( M_{i1} \) and \( M_{i2} \), \( i = 1, \ldots, n \), are the subject means (over the \( k \) replicate tests), then the estimated gradient is

\[
b = (S_{22} - \bar{S}_{11} + \sqrt{[(S_{22} - \bar{S}_{11})^2 + 4 \times 0 \times S_{12}^2]})/(2 \times S_{12})
\]

where \( S_{11} = \Sigma_{i=1}^n (M_{i1} - \bar{M})^2 \); \( S_{22} = \Sigma_{i=1}^n (M_{i2} - \bar{M})^2 \); and \( S_{12} = \Sigma_{i=1}^n (M_{i1} - \bar{M}) \times (M_{i2} - \bar{M}) \). \( M_1 \) and \( M_2 \) are the overall means for each test, i.e., \( M_1 = \Sigma_{i=1}^n M_{i1}/n \), \( h = 1, 2, \) and \( \theta = \sigma^2/\sigma_i^2 \).
The $\sigma_i^2$ and $\sigma_j^2$ are estimated from their respective $MS_w$, so that we estimate $\mu$ by

$$MS_w \text{ (test 2)/}MS_w \text{ (test 1)}$$

The intercept is then estimated as

$$a = M_2 - b \times M_1$$

This method is described and contrasted with other methods by Riggs et al. (27), where it is shown to perform well under a range of values of $i$ when the correlation between the $M_1$ and $M_2$ is fairly high (above ~0.5) and $i$ is estimated fairly precisely. Such conditions are likely to apply in these experiments; the $M_1$ and $M_2$ are measuring the same underlying physiological variable so the correlation will be high, and $\sigma_1$ and $\sigma_2$, and hence $\theta$, are directly estimated from the repeat measurements using each test.

**APPENDIX II: SIMULATIONS**

We used simulations to examine the distribution of the DR and loge(DR) and to check the accuracy of the Taylor series formula for the SD of loge(DR), given the form of model described by Eq. 1. These were performed for all combinations of the following values of $n$ (number of subjects), $k$ (number of replicate tests), and $\Delta$ (the true DR)

$$n = 5, 10, 15, 20, 30, 40, 60, 100, 200, 500$$

$$k = 2, 4, 6, 8, 10$$

$$\Delta = 1.5, 2.0, 3.0, 4.0, 5.0$$

For each of these combinations of $n$, $k$, and $\Delta$, the following procedure was performed.

1) An arbitrary subject mean $\mu$ was chosen, along with a set of $n$ equally spaced subject effects $\alpha_i$ chosen symmetrically around zero so that $\Sigma_{i=1}^{n} \alpha_i = 0$.

2) The $\sigma^2$, the within-subject variance, was calculated as

$$\sigma^2 = \left[ \sum_{i=1}^{n} \alpha_i^2 / (n - 1) \right] \left[ \Delta^2 \right]$$

3) For each $i = 1,...,n$ and $j = 1,...,k$, a random observation $e_{ij}$ was generated from a normal distribution with mean zero and variance $\sigma^2$. $X_{ij}$ were then generated from Eq. 1.

4) The DR and hence loge(DR) were calculated from the $X_{ij}$ using Eqs. 3–5.

5) The SD of loge(DR) was calculated from the Taylor series approximation (Eq. 17 of APPENDIX III), using the noncentrality parameter $\lambda$ evaluated from the DR estimate at the current step of the simulation.

6) Steps 3–5 were repeated 500 times, yielding a distribution of 500 values for each of DR, loge(DR), and SD of loge(DR).

7) The distributions of DR and loge(DR) were checked for normality using the Shapiro-Wilk test and were plotted as histograms.

8) The true SD of loge(DR) was estimated from the simulated distribution of loge(DR).

9) The distribution of Taylor series estimates of the SD of loge(DR) was compared with the true SD by plotting the median, upper and lower quartiles, and 5% and 95% against $n$ for different values of $k$ and $\Delta$.

Examination of $P$ values from the Shapiro-Wilk test showed some evidence that loge(DR) was not quite normally distributed (slightly >10% of the $P$ values examined were <0.05, but there was no apparent relationship between low $P$ values and $n$, $k$, or $\Delta$). However, this was a marked improvement over the DR itself, for which >50% of the $P$ values were <0.05. Histograms also showed the distribution of loge(DR) to be symmetric, whereas that of DR was markedly positively skewed (data not shown). Loge(DR) was deemed to be sufficiently close to normal for use in the $x^2$ test for equality of DRs. Plots showed that the median Taylor series estimate for the SD of loge(DR) was generally within ±10% of the true value for $n \geq 10$. However, for $k = 2$, the SD could be overestimated by as much as 20% for $10 \leq n \leq 20$. The distribution of SDs is positively skewed and, for $n \geq 10$, the 5% SD was at most 10% below the true value. Overestimation could be more marked, but even the upper quartile SDs were within ±25% of the true value (data not shown). Because the $x^2$ test statistic is a function of the reciprocal of the SD, the test is conservative with respect to overestimates of the SD, i.e., one is unlikely to wrongly reject the null hypothesis (no difference between the DRs), but the test may not be particularly sensitive to genuine differences for small values of $n$, especially if $k = 2$.

Estimates of the SD become very inaccurate for $n < 10$, and the approximation should not be used in this range. However, we would not recommend performing an evaluation study of this kind on such a small number of subjects, since the objective is to characterize the performance of the tests over a reasonable range of the variable of interest.

**APPENDIX III: SD OF LOG e(DR) USING TAYLOR SERIES APPROXIMATION**

We derived an estimate of the variance (and hence the SD) of loge(DR) using a first-order Taylor series expansion. From Eq. 7

$$DR = \sqrt{[(F_0 - 1)k]}$$

where $F_0 = MS_0/MS_w$ has a noncentral F distribution with degrees of freedom $v_1 = n - 1$ and $v_2 = n \times (k - 1)$ and noncentrality parameter

$$\lambda = k \times \sum_{i=1}^{n} \alpha_i^2 / \sigma^2 = (n - 1) \times k \times \Delta^2$$

Let $LDR = loge(DR)$

Expanding $LDR$ as a Taylor series in $F_0$ about its mean $F_0$, gives, to first order

$$LDR = LDR' + (F_0 - F_0') \times d(LDR)/dF_0$$

where $LDR'$ is $LDR$ evaluated at $F_0$ and $d(LDR)/dF_0$ is also evaluated at $F_0$. Hence

$$\text{var}(LDR) = \text{var}(F_0) \times [d(LDR)/dF_0]^2$$

where $\text{var}$ indicates variance. Now

$$d(LDR)/dF_0 = (1/DR') \times \frac{1}{2} \times [(F_0 - 1)/(k)^{1/2} \times (1/k)]$$

$$= 1/[2k \times (DR')^2]$$

where $DR'$ is $DR$ evaluated at $F_0$. From general properties of the noncentral F distribution (17)

$$F_0 = v_2 \times (v_1 + \lambda)/[v_1 \times (v_2 - 2)] = n \times (k - 1)$$

$$\times (n - 1 + \lambda)/(n - 1) \times (nk - n - 2)]$$

$$\Rightarrow DR' = [n \times (k - 1) \times \lambda + 2$$

$$\times (n - 1)]/[k \times (n - 1) \times (nk - n - 2)]$$
and \( \text{var}(F_0) = 2 \times (v_2/v_1)^2 \times [(v_1 + \lambda)^2 + (v_1 + 2\lambda) \times (v_2 - 2) \times (v_2 - 4)] \times (v_2 - 2)/(v_2 - 2) \times (v_2 - 4) \) 

\[
= 2n^2(\lambda^2 + \lambda \times (n - 3) + (n - 1)) \times (n - 3)/(n - 1)^2 \times (n - n - 2)^2
\]

Substituting for \( DR \) from Eq. 15 into Eq. 14 and for \( d(LDR)/dF_0 \) from Eq. 14 and \( \text{var}(F_0) \) from Eq. 16 into Eq. 13 gives

\[
\text{var}(LDR) = n^2(\lambda^2 + \lambda \times (n - 3) + (n - 1) \times (n - 3)) \times (2 \times (n - n - 4) \times [n(\lambda - 1) \times 2(\lambda - 1)]^2)
\]

which is evaluated by noting that

\[
\lambda = (n - 1) \times k \times \Delta^2
\]

and replacing \( \Delta^2 \) by \( DR^2 \), i.e.

\[
\lambda \approx (n - 1) \times k \times DR^2
\]

We are grateful for the assistance from Dr. Sue Manley and Nuala Walravens. This study was done with a grant from Servier and the Alan & Babette Sainsbury Trust.


Received 30 December 1997; accepted in final form 15 October 1998.

REFERENCES