Evidence-based medicine. Part 3: Diagnostic tests

Tim Holt

The diagnostic criteria for diabetes have long been a topic of debate. Recently, the American Diabetes Association and the World Health Organization recommended that HbA_{1c} be used as a diagnostic test, but UK organisations have yet to make a decision, as the methods for identifying a reliable, sensitive and specific test are complex. The third in a series looking at evidence-based medicine, this article looks at the methods of researching diagnostic tests, with HbA_{1c} as an example.

Diagnosing diabetes is an evolving process. The tasting of urine for sweetness was originally used in ancient Greece (Berger, 1999) but continued at least into the 19th century and was described by the frontier doctors of the American West (Dary, 2008). In the absence of laboratory blood glucose measurement, this finding added value to a clinical assessment, helping to distinguish between diabetes mellitus and diabetes insipidus, both characterised by polyuria and dehydration.

Much later, the change in diagnostic threshold for fasting plasma glucose from >7.8 mmol/L to >7.0 mmol/L in 1999 added an artificial boost to the genuinely escalating global diabetes prevalence of the late 20th century (World Health Organization [WHO], 1999). Still more recently, a move to base the diagnosis on HbA_{1c} level has been accepted by the American Diabetes Association (ADA, 2010) and by the WHO (2011).

With the increasing availability of diagnostic tests, the latter half of the 20th century witnessed another phenomenon: the increasing use of screening to facilitate early detection and intervention for a range of diseases. The criteria for an effective screening programme were first outlined for the WHO by Wilson and Jungner (1968).

A clinical diagnosis supported by laboratory evidence

Laboratory evidence of diabetes should be interpreted in the clinical context. Diabetes is a state of chronic hyperglycaemia. Raised glucose levels may complicate acute illnesses but return to normal after recovery, so a persistent tendency to hyperglycaemia should be evident before a diagnosis is made (WHO, 1999). This is one advantage of the HbA_{1c} approach, as its value depends on blood glucose levels over 2–3 months rather than single values that might be spuriously raised and atypical for the individual. Individuals will often have had diabetes for several years before diagnosis. Whichever approach is used, an abnormal result suggesting diabetes should

Article points

1. The diagnosis of diabetes continues to evolve with the use of modern laboratory tests.
2. The "sensitivity" of a test is the proportion of all people who have the condition that will show up as positive with the test."Specificity" indicates a test's ability to identify those unaffected.
3. To manage people cost-effectively, clinicians organising laboratory tests need to bear in mind the clinical picture, the likely pre-test probability in the individual, and the distinction between investigating established symptoms (based on clinical suspicion) and screening of low-risk populations.

Key words

- HbA_{1c}
- Receiver operator characteristic curve
- Sensitivity
- Specificity

Tim Holt is Academic Clinical Lecturer at the Department of Primary Care, University of Oxford, Oxford.

Page points

1. Very often a preliminary test (which is inexpensive and widely available) is used to identify individuals requiring confirmation through some "gold standard" measurement. This is the approach typically used in screening programmes.
2. The "sensitivity" of a test is the proportion of all people who have the condition that will show up as positive with the test (i.e. the true positives, divided by the true positives plus the false negatives).
3. "Specificity" indicates a test's ability to identify those unaffected. It is equal to the true negatives, divided by the true negatives plus the false positives.
4. The "area under the receiver operator characteristic curve" (AUROC) is given as a proportion of the maximum possible value of 1.0. A good test that performs well at identifying the condition will have a high AUROC, perhaps 0.85.

be confirmed by a repeated measurement in an asymptomatic person. This avoids incorrect diagnoses resulting either from laboratory error or self-limiting episodes of dysglycaemia. Certain conditions affecting red blood cell turnover can make HbA_{1c} measurements unreliable (for example, haemolytic states), as can iron deficiency. A clinical pattern suggestive of diabetes (polyuria, polydipsia, weight loss) greatly increases the significance of a single abnormal blood glucose value. Two abnormal values become more significant still. Conversely, a negative test for diabetes should not prevent us looking harder with repeated testing if clinical suspicion is high. One diagnostic test (such as HbA_{1c}) may identify a slightly different population than another (fasting plasma glucose or oral glucose tolerance test [OGTT]) (Manley et al, 2010).

This raises the issue of the ability to exclude or confirm important clinical states through a single diagnostic test. Very often a preliminary test (which is inexpensive and widely available) is used to identify individuals requiring confirmation through some "gold standard" measurement. This is the approach typically used in screening programmes.

Sensitivity and specificity

In diabetes, random blood glucose is sometimes used to identify individuals who may benefit from fasting plasma glucose or OGTT. Increasingly, HbA_{1c} is likely to be used in the initial stage although it is significantly more expensive. OGTT is inconvenient for patients and has resource implications, but will identify a higher proportion of people with diabetes than fasting plasma glucose alone. This means that it is more sensitive.

The "sensitivity" of a test is the proportion of all people who have the condition that will show up as positive with the test (i.e. the true positives, divided by the true positives plus the false negatives). If the false negative rate is low then the sensitivity will be high, and the test is good at identifying affected individuals. However, a test might show up positive for everyone in the population including those unaffected. It would then have 100% sensitivity but would be useless because it cannot identify those unaffected.

"Specificity" indicates a test's ability to identify those unaffected. It is equal to the true negatives, divided by the true negatives plus the false positives. A low false positive rate will give a high specificity, because a high proportion of the unaffected population are correctly identified.

A satisfactory trade-off indicates a useful test

As the threshold value for suggesting a diagnosis through a screening test is changed, the proportion of people correctly identified by the gold standard will change. Tests of high sensitivity may have low specificity, and vice versa. The trade-off between sensitivity and specificity can be plotted using a "Receiver Operating Characteristic" (ROC) curve (Zou et al, 2007, *Figure 1*). If the test is useless, the sensitivity will steadily increase in direct proportion to the reduction in specificity as the diagnostic threshold is varied. In the case of a useful test, the specificity will remain adequate as sensitivity increases up to an "optimal" value for both, and then falls off quite swiftly. The optimal value may be chosen for use in practice, although different thresholds may be appropriate according to whether we are interested in confirming a suspected diagnosis (investigating) or excluding a relatively unlikely diagnosis (as in screening).

This trade off can be measured by taking the "area under the ROC curve" (AUROC), which is given as a proportion of the maximum possible value of 1.0. A good test that performs well at identifying the condition will have a high AUROC, perhaps 0.85. A poor test will have a low AUROC, and in the case of a totally useless test it would be 0.5.

So the sensitivity and specificity of a diagnostic test vary according to the threshold value chosen. The AUROC measures the trade-off between them and the usefulness of the test in distinguishing cases from non-cases. It helps to identify the value of the test giving the optimal "compromise" between the two.

Background prevalence and pre-test probability

Unfortunately, understanding sensitivity and specificity is not the whole story. When we carry out a test and the result is positive, we would like to know what the chances are that the individual actually has the condition, i.e. how likely he or she is to show up as positive by the "gold standard" measure. What we are usually interested in is the positive predictive value (PPV), i.e. the probability that an individual testing positive actually has the condition. It might be assumed that this is also to do simply with the sensitivity and specificity of the test, but it also depends on the prevalence of the condition in the background population. If this is very low then even a test with high sensitivity and specificity may have a low PPV.

When we consult with patients and arrange blood tests, we are often confronted with a range of potential "boxes to tick" on the request form. In some cases we have a definite clinical suspicion that a disease is present (e.g. symptoms, signs, risk factors). In other cases, we have very little or no particular suspicion. In a sense this is similar to the distinction between investigating and screening, because even though patients in these two categories may come from the same practice population, they are drawn from different populations in statistical terms – populations with different prevalence values for the condition of interest. This prevalence is equivalent to the pre-test probability, and it determines the positive predictive value of the diagnostic test, even though the specificity and sensitivity are the same.

A good illustration of this comes from an article by Loong (2003), who provided a visually helpful means of explaining these terms. At the end of the article, the author describes the case of a person tested for anti-nuclear antibodies to detect systemic lupus erythematosus (SLE). The person has no clinical features of this condition, so the test is essentially being conducted as a screening test rather than an investigation. The background prevalence of SLE in the community is 30 per 100 000 and as there is no clinical suspicion, this is equivalent to the pre-test probability of SLE for this individual. Even though the sensitivity of the test is 94% and the specificity 97%, a positive test result will still only mean that the person has a 1% probability of

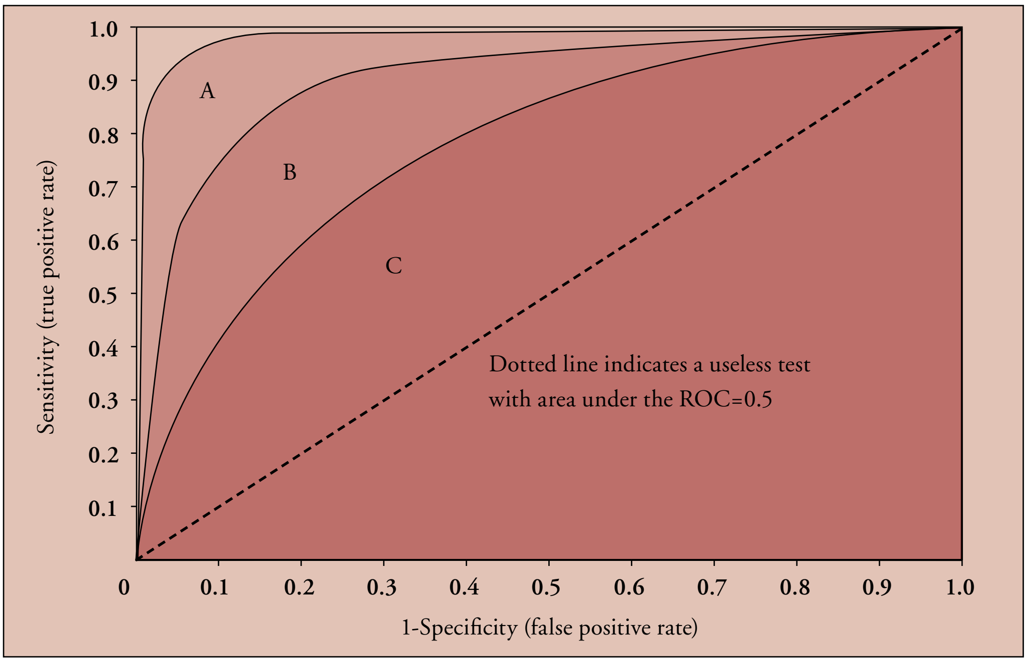

Figure 1. Three receiver operator characteristic (ROC) curves. Considering the area under the curve, test A is better at distinguishing cases from non-cases than B or C.

Page points

1. The positive predictive value is the probability that an individual testing positive actually has the condition.
2. The background prevalence in different populations determines the positive predictive value of the diagnostic test, even if the specificity and sensitivity are the same.

Page points

1. Whenever we are tempted to arrange a test "just to check" for some condition, we should consider what the pre-test probability is likely to be for the individual, based on overall clinical suspicion.
2. Population screening has often been advocated for diabetes, particularly in higher risk populations such as older people, those with high BMI, a family history, certain ethnic groups, and those with a history of gestational diabetes. But for each of these groups the value of testing asymptomatic people will depend on the pre-test probability in the individual, and this is very variable among different risk categories.
3. Lu et al (2010) adopted a two threshold approach to the diagnosis of diabetes using HbA_{1c}, as they recognised that using a single diagnostic "cut-off" for HbA_{1c} is limited by the area under the receiver operator curve for this test and for this condition.

having SLE. However, a different person with arthralgia, malar rash and proteinuria that tests positive has a much higher PPV because they come from a different population with a much higher SLE prevalence.

So whenever we are tempted to arrange a test "just to check" for some condition, we should consider what the pre-test probability is likely to be for the individual, based on overall clinical suspicion. These days it is all too easy to arrange tests without considering their value in the particular situation. Perhaps if we still had to taste urine we would think harder about whether we really did suspect diabetes. On the other hand, a low background prevalence raises the negative predictive value of the test, i.e. the probability that the diagnosis really is absent if the test is negative, and this may be justifiable in population screening particularly if the test is inexpensive.

Population screening has often been advocated for diabetes based on the Wilson and Jungner criteria (Engelgau et al, 2000), particularly in higher risk populations such as older people, those with high BMI, a family history, certain ethnic groups, and those with a history of gestational diabetes. But for each of these groups the value of testing asymptomatic people will depend on the pre-test probability in the individual, and this is very variable among different risk categories. Given the large numbers of people for potential inclusion in the community, it is not surprising that this issue has received much debate (Stephens and Williams, 2006). The result is that formal screening is currently only recommended by NICE for those undergoing cardiovascular risk assessment (NICE, 2008a), and for those with a history of gestational diabetes (NICE, 2008b).

The adoption of HbA_{1c} by the ADA, if accepted in the UK, is likely to make us tick the HbA_{1c} box rather than the random glucose box to "rule out" as well as positively diagnose diabetes. This may be very appropriate depending on the clinical situation but is significantly more expensive. The health economic case would need to be confirmed before, for instance recommending it for asymptomatic people with hypertension as part of a routine annual review, in the way that random blood glucose tests are often carried out now.

How would HbA_{1c} perform as a screening test?

In 2010 Lu et al examined the usefulness of HbA_{1c} both at excluding and diagnosing diabetes (Stephens and Williams, 2006). They identified a total of 8505 people from two separate sources (the Melbourne Pathology group and the AusDiab population), for which values of both HbA_{1c} and oral glucose tolerance tests were available. The Melbourne Pathology group was a population referred by general practitioners for OGTT, and had an overall prevalence of diabetes (by OGTT criteria) of 34.6%. The AusDiab group was drawn from a national population-based study, and the prevalence was much lower at just 4.6%. For different thresholds of HbA_{1c} they reported the proportion that would be diagnosed with diabetes by the "gold standard" OGTT.

The authors adopted a two threshold approach, as they recognised that using a single diagnostic "cut-off" for HbA_{1c} is limited by the AUROC curve for this test and for this condition. Even in their Melbourne Pathology group population with high diabetes prevalence, an HbA_{1c} level of 6.2% (44 mmol/mol) (identified as optimal by the ROC curve) produced reasonable values for sensitivity (82.2%) and specificity (78.8%), but a PPV of only 67.2%. By using two "cut-offs", the authors were able to demonstrate that a value of <5.5% (<37 mmol/mol) could be used to effectively rule out diabetes and a threshold of >7.0% (>53 mmol/mol) to rule-in the condition in both populations (and by implication in both investigation and screening scenarios). While values of 6.5–6.9% (48–52 mmol/mol) were likely to indicate diabetes, the specificity was lower and the PPV more dependent on the background prevalence.

Conclusion

The diagnosis of diabetes continues to evolve with the use of modern laboratory tests.

Interpreting test results requires an awareness of the particular clinical situation from which the test is drawn. HbA_{1c} has been accepted by the ADA as a means of diagnosing diabetes, and the UK may follow this policy in the foreseeable future. HbA_{1c} offers the advantage that it reflects blood glucose values over 2–3 months, but must be interpreted with caution in certain clinical states including haemolysis and iron deficiency.

No test is perfect and alternative diagnostic tests may have different sensitivity and specificity values depending on the source population. To manage people cost-effectively, clinicians organising laboratory tests need to bear in mind the clinical picture, the likely pre-test probability in the individual, and the distinction between investigating established symptoms (based on clinical suspicion) and screening of low-risk populations. ■

American Diabetes Association (2010) *Diabetes Care* **33**(Suppl 1): S62–9

Berger D (1999) *MLO Med Lab Obs* **31**: 28–30, 32, 34–40

Dary D (2008) *Frontier Medicine. From the Atlantic to the Pacific, 1492-1941.* Alfred A Knopf (Random House), New York

Engelgau MM, Narayan KM, Herman WH et al (2000) *Diabetes Care* **23**: 1563–80

Loong TW (2003) *BMJ* **327**: 716–19*

Lu ZX, Walker KZ, O'Dea K et al (2010) *Diabetes Care* **33**: 817–19

Manley S, Nightingale P, Sratton I et al (2010) *Diabetes & Primary Care* **12**: 87–96

NICE (2008a) *Lipid Modification: Cardiovascular Risk Assessment and the Modification of Blood Lipids for the Primary and Secondary Prevention of Cardiovascular Disease.* CG67. NICE, London

NICE (2008b) *Diabetes in Pregnancy: Management of Diabetes and its Complications from Pre-conception to the Postnatal Period. CG63.* NICE, London

Stephens JW, Williams R (2006) *Diabet Med* **23**: 1163–4

Wilson JMG, Jungner G (1968) *Principles and Practice of Screening for Disease.* World Health Organization, Geneva

World Health Organization (1999) *Definition, Diagnosis and Classification of Diabetes Mellitus. Report of a WHO Consultation.* WHO, Geneva

World Health Organization (2011) *Use of Glycated Haemoglobin (HbA_{1c}) in the Diagnosis of Diabetes Mellitus. Abbreviated Report of a WHO Consultation.* WHO, Geneva. Available at: http://bit.ly/fRfBnX (accessed 26.05.11)

Zou KH, O'Malley AJ, Mauri L (2007) *Circulation* **115**: 654–7

"To manage people cost-effectively, clinicians organising laboratory tests need to bear in mind the distinction between investigating established symptoms (based on clinical suspicion) and screening of low-risk populations."

*Warning: this is a classic article but contains some errors, including the subtitle. These are all highlighted in the rapid responses.