Testing challenges: evaluation of novel diagnostics and molecular biomarkers

Ronald L Zimmern

This article is based on the Milroy Lecture given at the Royal College of Physicians on 7 May 2008 by **Ronald L Zimmern** MA FRCP FFPHM, Executive Director, PHG Foundation, Cambridge; Consultant in Public Health Medicine, Cambridge University Hospitals NHS Trust; Associate Lecturer, University of Cambridge

Clin Med 2009;9:68–73

ABSTRACT – Through the lens of public health genomics, this article probes certain issues that concern the evaluation of diagnostic tests and molecular biomarkers, and the accompanying policy and regulatory implications. It begins with some conceptual remarks followed by a discussion of evaluation, translation, and regulation, and their importance for public health.

KEY WORDS: biomarkers, diagnostic tests, translation, regulation

Conceptual issues

Diagnosis is about classification and how it may be used as a label to aid prediction, prognosis and treatment.[1] The label, the diagnosis, is not an end in itself but an intermediary, a means to an end. Diagnosis is no use in itself; there must be a purpose, an objective. Tests, including the use of clinical symptoms and signs, are the means by which a diagnosis is made so that a decision or an action can be taken. It is also the case that one can make no statement about the effectiveness of a test without knowing its purpose or objective since purpose is inherent in the formal definition of the effectiveness of a healthcare intervention. But it is not just purpose that is important in test evaluation. The nature of the disease is also important. The effectiveness, validity or utility of a test is dependent on the disease or disorder under consideration. The third factor that has a bearing on test interpretation and evaluation is population and, in particular, the effect that the disease prevalence in the studied population critically affects the test's predictive value.[2]

The term 'biomarker' is often used in this context rather than diagnostics or diagnostic tests. It has been more broadly defined by the Food and Drug Administration as a characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention. There are many reasons for carrying out a test, of which the making of a diagnosis is but one (Fig 1).

Fig 1. Test purposes.

Purpose or uses of a test or biomarker

1 Diagnosis
2 Disease classification
3 Risk stratification
4 Disease prognosis
5 Treatment stratification
6 Treatment monitoring
7 Population screening

These considerations lead to the most important conceptual insight into biomarker or diagnostics evaluation, the distinction between an 'assay' and a 'test'. The assay is a method for determining the presence or quantity of a component whereas the test is its use in the context of a particular disease, in a particular population, for a particular purpose. The distinction has an important practical implication. Whereas the evaluation of an assay is reasonably straightforward and allows broadly applicable standards to be established, the evaluation of a test is more complex and inherently less susceptible to standardisation. Each test is likely to need evaluation and interpretation depending on how the test is to be used in the particular context of disorder, population and purpose.[2]

Evaluation

Various frameworks have been developed for this process. The genomics community has settled on the ACCE framework to guide its activities (Fig 2).[3] The analytical validity is essentially a measure of the technical evaluation of the assay used in the test. It defines the ability of the assay to measure accurately and reliably the component of interest. Its performance is judged against an agreed reference standard.

The clinical validity of a test by contrast defines the ability of the assay to detect or predict the presence or absence of clinical disease or risk of disease in the context of population and purpose. It is of the utmost importance to understand that there are two separate aspects to clinical validity. First, there is the need to show evidence of biomarker-disease association. This is essentially a matter for epidemiological studies which are normally carried out by the scientific community. Second, assuming that such an association has been demonstrated and validated, it will additionally be necessary to ascertain test performance using measures such as sensitivity, specificity, positive and negative predictive values, likelihood ratios (both positive and negative), and area under the receiver-operator curve (ROC). The fact that there is well confirmed biomarker-disease association does not entail that the performance of the test is necessarily valid, that it necessarily serves to discriminate accurately between an individual with or at risk of a disease from one who is without disease or at lower risk. Validity cannot be assumed; it must be empirically determined.

Clinical utility refers to the likelihood that the test will lead to an improved outcome. It is here that the purpose of the test becomes an essential element. To date, this has been a much neglected area, but recently ways by which to set out dimensions of clinical utility have been explored.[4]

This analysis holds true for diagnostic tests, but its application to predictive tests will need some modification. Whether predictive or diagnostic the actual reality is that for many tests in use formal evaluations along the lines suggested above have not been carried out. It has been proposed that for simpler tests, such as serum sodium or a white blood count, such formality is not required. The failure to evaluate the newer complex molecular biomarkers will be highly detrimental for the proper care of patients and for the financial health of the NHS.

Translation

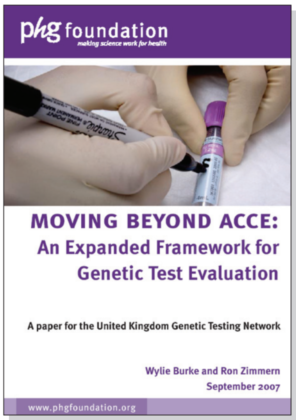
This section discusses the translation of scientific advances, in particular biomarkers and diagnostic tests, into effective technologies and interventions for individuals and populations. The Cooksey Report, published in December 2006, set out a strategy for the funding of UK health research.[5] It outlines a pathway from basic research through preclinical development, clinical trials, health technology assessment and health services research into healthcare delivery. It identified two gaps in translation: the first arising from the translation of the results of basic scientific or clinical research into products that might, in due course, be disseminated into wider healthcare practice – a process that has been summarised in the phrase 'from bench to bedside' and called by others 'type I translation'; the second relating to the evaluation of these new products and their implementation into routine clinical practice – 'research into practice' referred to by some as 'type II translation'. Both gaps, it suggested, needed to be addressed through translational research.

These ideas are, of course, not new. The Shattuck Lecture in 2003 in the *New England Journal of Medicine* was entitled 'Clinical research to clinical practice – lost in translation?'. It bemoaned the fact that it took far too long before clinical practice caught up with the results of research findings.[6] Elias Zerhouni, Director of the National Institutes of Health (NIH), proposed a new vision in 2003, the NIH Roadmap, in which he stated that it was necessary to 're-engineer the clinical research enterprise' such that 'translational and clinical research are core components of a full-spectrum biomedical research enterprise'.[7] Despite all this, the emphasis and research funding continues to be directed at basic research and at type I translation, as evidenced by research on publications in the genetic literature, which showed that under 3% was on research that fell outside the range of type 1 translational activity. Indeed Lawrence Green from the University of California dubbed type II translation as 'the roadmap less travelled'.[8]

The other point that might be made about Cooksey and translation is that it failed, first to mention the importance of the population sciences, especially biostatistics and epidemiology, in the translation process; second, that there was a social, legal, ethical and policy context to consider; and third, that outcomes could not be confined to service delivery but had to include policy development.

Jonathan Lomas was quite explicit in stating, 'For research findings to effectively influence health services' delivery of care needs an 'intermediary'.[9] This intermediary is the knowledge broker; in Lomas' view, an entrepreneurial, trusted, credible communicator who understands the cultures of the research and decision-making environments and is able to facilitate, mediate and negotiate the way between the one and the other culture. Both research and decision making are complex processes, not products or events, which need active management by an agent who is able to interpret between the two cultures in order to drive the change necessary to get research into practice. The literature on translational research has not attended adequately to the concepts in Lomas' paper, and policy makers have on the whole failed to distinguish (and have even conflated)

Fig 2. The ACCE Framework.



phg foundation
making science work for health

MOVING BEYOND ACCE:
An Expanded Framework for
Genetic Test Evaluation

A paper for the United Kingdom Genetic Testing Network

Wylie Burke and Ron Zimmern
September 2007

www.phgfoundation.org

1 **A** nalytical validity
2 **C** linical validity
3 **C** linical utility
4 **E** thical, legal and social

Analytical validity of a test defines its ability to measure accurately and reliably the component of interest

Clinical validity of a test defines its ability to detect or predict the presence or absence of clinical disease or predisposition to disease

Clinical utility of a test refers to the likelihood that the test will lead to an improved outcome

Ethical, legal and social implications of a test

The ACCE framework is applicable to all forms of molecular diagnostics and biomarkers

translational research from the process of translation as envisaged by Lomas. This failure (seen also in Cooksey), to distinguish and to adequately fund, is particularly starkly seen when considering diagnostics and biomarkers.

Regulation

Regulation is a term that is often used synonymously with statutory regulation – in effect, a legal restriction promulgated by government and supported by a threat of sanction or a fine. This, however, is far too narrow and the word should be used (at least in the context of public policy) to encompass any form of control or governance of behaviour. It has been suggested that in the context of genetic tests, essentially their regulation could be thought of as taking place in three separate domains – through statutory mechanism, through control of resources by funders or reimbursers of healthcare, and through mechanisms of clinical governance at the level where physicians and patients interact.[10] Over the past three years, working with David Melzer at the University of Exeter on a project funded by the Wellcome Trust, the policy issues in genetic testing have been explored. Key players in this field on both sides of the Atlantic were interviewd, individually and in focus groups. The conclusions reached in this study were echoed and reinforced by the deliberations of a diagnostic summit organised by the PHG Foundation and the Royal College of Pathologists in January 2008. A summary of its key recommendations are shown here (Fig 3).[11,12]

Statutory regulation

Biomarkers are regulated as medical devices and, in the UK, this is carried out by the Medicine and Healthcare products Regulatory Agency (MHRA). The regulatory framework is based upon the European *In Vitro* Diagnostic Medical Devices Directive 98/79/EC. The language that is used suggests that it is there to regulate tests and diagnostics but, if you accept the earlier distinction between an assay and a test, it can be suggested that de facto the MHRA (and other regulatory agencies) function primarily to regulate the integrity of the assay and to ensure its safety as a product. The question is whether the regulator should go beyond this brief? To what extent should it demand the provision of the evidence base for clinical validity or utility and prevent an assay from being marketed without such evidence? What role should pre-market-review play? Put more starkly, should the regulator be concerned with products that are safe but ineffective?

Additionally, there is the question of what is meant by ‘safe’ when dealing with biomarkers? This question is much easier to answer for drugs and other therapeutic interventions where matters of toxicity or adverse reactions can be clearly delineated. It is far from clear when considering biomarkers.

Statutory regulation is not an appropriate vehicle for regulating the clinical performance of tests and biomarkers; it should, as at present, confine itself to the technical regulation of the assay. The one exception, perhaps, is that where clinical claims are made, the regulator should require evidence of a true and validated biomarker-disease relationship. Yet in so concluding, we are left with a dilemma. How should we ensure that only those biomarkers that have evidence of effectiveness be used in practice?

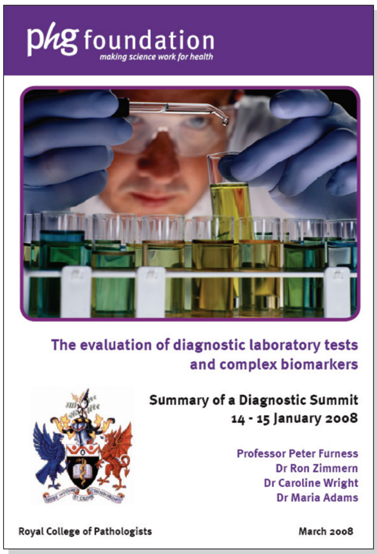
There is much that can be said about this matter. But in brief, first, some reliance has to be placed on the other two domains of regulation – reimbursement and clinical governance; and second, the mechanism that will allow these domains to be effective in their task is the establishment of an open and transparent database of evidence.

Public health implications

It is now no longer acceptable to take the view that physicians can, purely as a result of their training, interpret all test results effectively without some help. One reason is that commissioners and funders of health services throughout the developed world are under extreme financial pressure and will require evidence of effectiveness before funding a test. This is essentially the reason why an evidence base is necessary. A second is that tests are more complex, being made more generally available, and have been increasing in numbers. I suggest also that tests that predict the risk of disease, or allow the prognosis of disease to be determined, will be of increasing importance in the coming decades.

Although the NHS keeps very little data on tests and test costs, it has been estimated by the Audit Commission that the total number increased by 6.4% per year (in biochemistry and haematology) and 9.5% per year (in

Fig 3. Summary of recommendations of diagnostic summit.



phg foundation
making science work for health

The evaluation of diagnostic laboratory tests and complex biomarkers

Summary of a Diagnostic Summit
14 - 15 January 2008

Professor Peter Furness
Dr Ron Zimmern
Dr Caroline Wright
Dr Maria Adams

Royal College of Pathologists
March 2008

1 A new body should be established to ensure the valuation of diagnostic tests.

2 A publicly available database be created of new and existing laboratory tests – a ‘diagnostics formulary’ – containing evidence for clinical performance, and explicitly stating where any evidence is lacking.

3 Policy makers and industry should be encouraged to address issues around gathering the necessary evidence for clinical evaluation.

4 An independent expert body should be responsible for evaluating the evidence for test performance and for making recommendations about appropriate clinical use.

5 Commissioners and healthcare professionals should be encouraged to use only those tests where appropriate evidence of clinical performance exists.

6 Statutory regulators should be empowered to require transparency relating to evidence of test performance, and ensure responsive and proportionate risk assessment to ensure patient safety.

microbiology) between 2000–1 and 2005–6. In East Anglia there has been a 14% per year compound growth in molecular genetic testing between 2002–3 and 2007–8. These numbers, between 2,500 and 6,000, can be put into context. Addenbrookes Hospital carried out 3,252,590 pathology tests in the 2007–8 financial year of which 1,921,273 (59.1%) were for biochemical tests. In financial terms, pathology accounted for £18m (4.97%) of the trust's £362m expenditure.[13]

A third reason is that predictive tests bring their own challenge and should be handled differently to conventional diagnostics. There is a real sense that predictive (pre-disease) biomarkers differ from post-disease biomarkers where, in effect, the disease is already present. The diagnostic biomarker, whether radiological, biochemical or physiological, serves as a consequential indicator of the disease process. The predictive biomarker, by contrast, is an indicator of future disease; disease which, by definition, is not present and where the biomarker is not a consequence of disease. It is a risk factor which may either be on the causative pathway to disease or is linked with some other factor that is on that pathway. The standard performance characteristics used for diagnostic tests, such as sensitivity or specificity and the dichotomisation of test results into positive or negative may not provide the correct approach for conceptualising how such predictive tests should be evaluated and used.

The new genetic and molecular biomarkers, unlike conventional diagnostics, also tend to be characterised by low relative risks (RRs) or odds ratios (OR). Effect sizes are usually less than 1.5 and often in the range of 1.05 to 1.2. These are certainly the sorts of numbers that are seen emanating from whole genome scans in various complex disorders. Sceptics have correctly pointed out that the low RRs of 1.0 to 2.0 will never allow one, using a single marker, to distinguish (with acceptable levels of false positives and negatives) those who will develop disease from those who will not.[14] By contrast, multiple biomarkers may have some utility.

An example is provided from the field of prostate cancer where the RR of individuals with a family history and one abnormal allele is 1.62, whereas with five or more abnormal alleles the risk goes up to 9.46.[15] A further example is provided by considering the genetic variant TCF7L2 and type 2 diabetes. The OR of the heterozygous carrier in one study was shown to be 1.35 and of homozygotes 1.90. But only 32% of homozygotes actually develop diabetes as compared to 23% in those with the wild type alleles, an absolute risk difference of only 9%.[16] The key question for clinicians is how these data should be used in a clinical or public health setting, and in particular to understand the absolute risks at which a patient or population group might appropriately (with benefit) undergo an intervention. Many in the UK, for example, might not recommend prostrate-specific antigen testing in an asymptomatic male, but would that advice hold for those who have all five of the abnormal alleles in the example cited above? The answer is unknown.

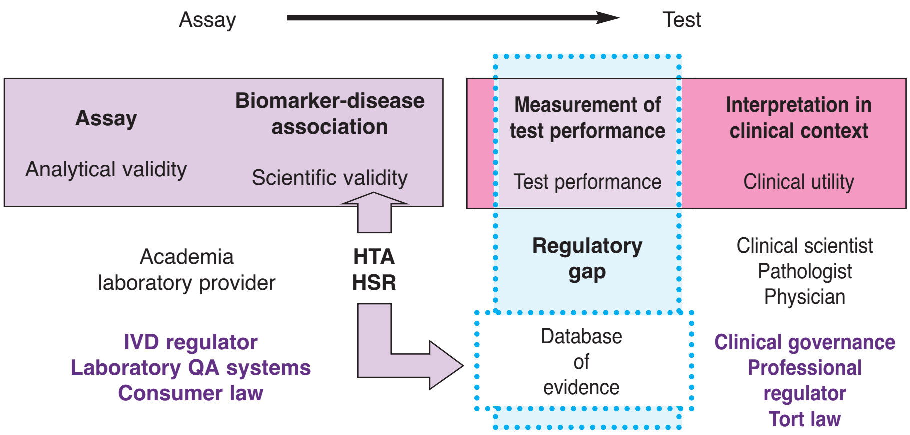
Direct consumer testing

There has been much anxiety over their lack of regulation and the fact that these tests are offered to the public with a very poor evidence base. Caroline Wright has reviewed 29 companies and the level of information that they provide against guidelines for direct to consumer testing from the American Society of Human Genetics. Wright has shown that none provide evidence of clinical validity and few point out the risks. Scientific publications are only cited in 50% of cases. At an epidemiological level, Cecile Janssens and colleagues have studied in detail the gene-disease association of tests offered by seven companies directly to consumers. The salient result from their study was that significant associations were only found in 38% of those polymorphism-disease associations that they investigated.[17]

A regulatory gap does exist but I am not convinced that the statutory regulators of *in vitro* devices such as the MHRA can easily carry out this function on their own. To the extent that claims are made for biomarker-disease association by test manufacturers, it would not seem unreasonable for the regulators to require evidence of scientific validity, namely that such claims are substantiated in the scientific literature – since by definition a false or unsubstantiated association cannot result in a clinically useful test. But once an association is substantiated, no matter how small, it would seem that there is a small chance that the test may be clinically valid or useful.

The question of how and where the responsibility for regulation should lie may be discussed in the light of the distinction between the assay as the measurement, and the test as the interpretation of that measurement. It has already been noted that the assay is the responsibility of the statutory regulator of *in vitro* devices. Interpretation, by contrast, has always been the responsibility of the clinician, and where it is the interpretation of the

Fig 4. Regulation of biomarkers. HSR = health services research; HTA = health technology assessment; IVD = *in vitro* diagnostic.



implications of a test result, as distinct from the measurement, that causes a problem, professional regulation rather than device regulation should be used in the solution (Fig 4).

Implications to health promotion and disease prevention

As a consequence of the growth in complex molecular biomarkers, a shift in the paradigm in terms of how health is promoted and disease is prevented may need to be considered. Public health interventions have classically been directed at the external environment – matters such as water or air quality – or to economic, political and social factors, such as fiscal policy in relation to alcohol which affect whole populations and communities. In recent years, attention has been additionally directed at behavioural determinants of health. By so doing the idea of a population has been altered in a subtly different way – conceptualising it not as a single entity but as collections of individuals.[18]

The development of genetic and molecular biomarkers allows us to differentiate individuals within populations even further by categorising them into groups based on estimates of their absolute risk of disease. The preventive intervention would not be uniform across a population but would vary according to an individual's risk. Cardiovascular risk assessment based on age, sex, cholesterol levels, blood pressure (BP) and smoking history provides an example of this type of preventive intervention that is already being used in general practice. Although all three approaches may be characterised as preventive interventions they have different social, legal and ethical implications (all of which need further exploration), and with the progress in genetic, cell and molecular science, the move towards such individualised or stratified prevention will accelerate – with profound implications for the practice of public health and health promotion.

Margaret Pepe suggested that by comparing the probability density function of a risk factor among persons who will develop the disease and those who will not, biomarkers of low RRs necessarily fail to meet the standards for a credible and valid test. In the case of preventive biomarkers, this may not be so. Indeed, raised cholesterol or BP levels that are used routinely in practice probably confer ORs of between 2 and 10, depending on the magnitude of the rise.[19]

Paul Pharoah uses as an example a scenario that once the main genetic determinants of a disease, such as breast cancer, have been elucidated, it will be possible, for example, to divide the female population into different groups according to genetic or biological risk; and to give different advice about mammography according to where the individual sits on this continuum.[20]

The generalisation of this insight is that more attention should perhaps be paid to an understanding of absolute risk and the levels at which individuals undergo particular interventions for certain diseases. The critical factor then becomes not so much the RR that the possession of a biomarker will impose (since with most the ORs will be relatively low) but an understanding of the baseline absolute risk of the individual and the absolute risk threshold for a particular intervention. This approach accords closely with how clinicians take decisions, and the notion that all tests are done for a particular purpose and with some action in mind.

Using this notion, predictive biomarkers may be viewed as genetic variants with low ORs, not as a diagnostic test in the conventional sense, but as extra pieces of data that will serve to modify absolute risk estimations. On this interpretation, all biomarkers, no matter how small the effect size, provided the disease association has been shown to be real, may have a potential use.

Conclusion

1. The distinction between an 'assay' and a 'test' is crucial to understanding the roles of statutory regulators and others in test evaluation.
2. Clinical validity requires more than evidence of gene-disease association; it also requires evidence of test performance such as sensitivity, specificity, positive and negative predictive values.
3. The main problem is lack of data; policy is urgently needed to establish systems and resources to generate evidence of test performance, and to agree the respective roles and responsibilities of government, statutory regulators, public bodies, academia and the commercial sector.
4. Systems should be established to ensure that the data are appropriately analysed and evaluated against agreed standards and that the evidence is placed in the public domain.
5. Funders and reimbursers of health services and clinicians should be discouraged from purchasing and using tests that are not backed by appropriate clinical evidence.
6. The role of statutory regulators should be confined to (a) ensuring the safety of all tests and biomarkers (b) ensuring that claims for biomarker-disease associations are genuine and real, and (c) requiring all evidence of test performance (or lack of it) to be placed in the public domain.
7. The public health community should anticipate the advent of stratified prediction and prevention and its implications for health promotion messages.
8. The conventional model for diagnostic tests may not be appropriate for predictive biomarkers. Instead attention should be directed at understanding absolute risk and how biomarkers may alter such risk in the context of preventive interventions.

Acknowledgements

I should like to thank the following for many hours of informal discussion and chat about diagnostic testing: Hilary Burton, Wylie Burke, Doug Easton, Peter Furness, Stuart Hogarth, Mark Kroese, Kathy Liddell, David Melzer, Christine Patch, Paul Pharoah, Chris Price, Simon Sandersona and Caroline Wright. Conversations with them did much to shape my thoughts.

References

1 Price CP, Christenson RH. The clinical question; a system for formulating answerable questions in laboratory medicine. In: Price CP, Christenson RH (eds), *Evidence based laboratory medicine. Principles, practice and outcomes*. Washington DC: AACC Press, 2007: 25–52.

2 Zimmern RL, Kroese M. The evaluation of genetic tests. *J Pub Health (Oxf)* 2007;24:1–5.

3 Kroese M, Zimmern R, Sanderson S. Genetic tests and their evaluation: Can we answer the key questions? *Genet Med* 2004;6: 475–80.

4 Burke W, Zimmern R. *Moving beyond ACCE: an expanded framework for genetic test evaluation*, 2007. www.phgfoundation.org/file/3736

5 Cooksey D. *A review of UK health research funding*. Norwich: Stationery Office, 2006. www.official-documents.gov.uk/document/other/0118404881/0118404881.asp

6 Lenfant C. Shattuck Lecture. Clinical research to clinical practice – lost in translation? *N Engl J Med* 2003;349:868–74.

7 Zerhouni EA. Translational and clinical science – time for a new vision. *N Engl J Med* 2005;353:1621–3.

8 Green LW. Translation 2 research. The roadmap less travelled. *Am J Prev Med* 2007;33:137–8.

9 Lomas J. The in-between world of knowledge brokering. *BMJ* 2007; 334:129–32.

10 Burke W, Zimmern R. Ensuring the appropriate use of genetic tests. *Nature Rev Genet* 2004;5:955–8.

11 Melzer D, Hogarth S, Liddell K, Ling T, Sanderson S, Zimmern R. *Evidence and evaluation: building public trust in genetic tests for common diseases*, 2008). www.phgu.org.uk/file_gateway?link_ID=4003

12 Furness P, Zimmern R, Wright C, Adams M. *The evaluation of diagnostic laboratory tests and complex biomarkers*, 2008. www.phgu.org.uk/file_gateway?link_ID=3998

13 Carr C, Kroese M, Whittaker J. Personal communication, 2008.

14 Holtzmann NA, Marteau TM. Will genetics revolutionize medicine? *N Engl J Med* 2000;343:141–4.

15 Zheng SL, Sun J, Wiklund F *et al*. Cumulative association of five genetic variants with prostate cancer. *N Engl J Med* 2008;358:910–9.

16 Groves CJ, Zeggini E, Minton J *et al*. Association analysis of 6,736 UK subjects provides replication and confirms TCF7L2 as a type 2 diabetes susceptibility gene with a substantial effect on individual risk. *Diabetes* 2006;55:2640–4.

17 Janssens AC, Gwinn M, Bradley LA *et al*. A critical appraisal of the scientific basis of commercial genomic profiles used to assess health risks and personalize health interventions *Am J Hum Genet* 2008; 82:593–9.

18 Sober E. Evolution, population thinking, and essentialism. *Philos Sci* 1980;47:350-83.

19 Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol* 2004;159:882–90.

20 Pharoah PDP, Antoniou AC, Easton DF, Ponder BAJ. Polygenes, risk prediction, and targeted prevention of breast cancer. *N Engl J Med* 2008;358:45–52.