Risk scores for predicting type 2 diabetes: comparing axes and spades

N. J. Wareham · S. J. Griffin

Received: 17 January 2011 / Accepted: 19 January 2011 / Published online: 5 March 2011
© Springer-Verlag 2011

Keywords Prediction · Risk scores · Type 2 diabetes

Abbreviations

DETECT-2	Evaluation of Screening and Early Detection Strategies for Type 2 Diabetes and Impaired Glucose Tolerance
ROC	Receiver operating characteristic

A recent systematic review of risk prediction tools for identifying individuals at risk for type 2 diabetes identified a total of 46 different studies reporting the derivation of risk models [1]. These scores included diverse types of variables ranging from simple information generally available in health records through to more detailed biochemical and genetic variables. The paper by the Evaluation of Screening and Early Detection Strategies for Type 2 Diabetes and Impaired Glucose Tolerance (DETECT-2) group in this issue of *Diabetologia* reports the results of an evaluation of a modified version of one of the most widely investigated diabetes risk prediction tools, the Finnish Diabetes risk score [2]. In this paper the variables included in the prediction model were age, BMI, waist circumference, use of blood pressure medication and previous high blood glucose defined as history of gestational diabetes. The authors compare the predictive ability of this new score with that of the original Finnish risk score and a number of other models. The DETECT-2 group's conclusion was that

this new model performed better than existing risk scores as assessed by the area under the receiver operating characteristic (ROC) curve.

The area under the ROC curve is a summary measure of predictive ability. However, consideration of how such a risk score might be used is necessary to truly determine its predictive ability. The modified Finnish Risk questionnaire is designed to be sent out to a target population to assess risk to determine who should be invited for more detailed biochemical testing. If a cut-off point of seven or greater were chosen, then 40% of the population who completed the questionnaire would be deemed to be 'screen positive' and would need to be invited for subsequent testing. It is true that with this cut-off point the sensitivity of the risk model is 76% or, in other words, the test identifies three-quarters of the people who will develop diabetes. However, high sensitivity is not possible without low specificity—there is always a trade-off. In this case, the relatively low specificity of 63% means that, for every nine people invited for the more detailed testing, only one will become a case over the next 10 years. Since such a simple screening tool is designed to be used in the context of a public health approach to the identification of risk groups, the need to test 40% of the population represents a logistical and financial challenge. Of course, a different cut-off point could be chosen, resulting in fewer people receiving an invitation to a follow-up test, but then the sensitivity of the test would be lower. Since there is always a trade-off, how does one choose the most appropriate cut-off point? There are statistical approaches for this but the true answer is that it depends upon the clinical implications of being a false negative compared with a false positive. In a situation where the consequences of being a false positive are high compared with those of being missed as a case (false

N. J. Wareham (✉) · S. J. Griffin
MRC Epidemiology Unit, Institute of Metabolic Science,
Addenbrooke's Hospital,
Cambridge CB1 0QQ, UK
e-mail: nick.wareham@mrc-epid.cam.ac.uk

negative), then one would always tend to optimise specificity over sensitivity. The reverse would be true in a situation in which a disease was serious but easily treated and the consequences of being falsely labelled were relatively minor. In this scenario, one would seek to optimise sensitivity. This way of thinking has not been applied often enough in the diabetes world and is preferable to the simple presentation of the area under the ROC curve since it begins to contextualise risk scores to the situation in which they will be used and the clinical or public health purpose that they are intended to serve.

Since these risk scores are being proposed for real clinical situations, one must view them in that light. In the analysis of the modified Finnish risk score, the DETECT-2 group compare prediction using data obtained in five prospective cohort studies. Research studies tend to have standardised and complete baseline measurements for their cohort. This is not the case in real life, and missing and inaccurate data would lower the predictive ability of such a score. If we imagine the use of such a risk questionnaire in a population setting, then one would first identify the target population from a population sampling frame and then post out questionnaires for the people identified to complete and return. However, perhaps only half the population would return the questionnaire. The analysis of predictive ability in the DETECT-2 paper overlooks this non-response. This is important if one is comparing risk score methods that have different response rates. Risk models like the Cambridge Risk Score [3] or the QDScore [4] use information already available in GP records and do not require the collection of new data from participants. Therefore, the response rate is closer to 100% in a setting such as the UK where population registers are based on primary care data and basic clinical information is relatively complete. If the Finnish risk questionnaire was posted out and received a response rate of 50%, then the true sensitivity at a cut-off point of seven or higher would not be 76% but, rather, 38%, and the questionnaire's predictive ability would not look quite so good when compared with the other scores that are designed to be implemented in a fundamentally different way. It is also likely that non-response will not be random but will probably be biased towards completion of the questionnaire by the healthier and socially advantaged in society, potentially missing many of the key subgroups that one truly wishes to target in a public health risk assessment programme.

Of course, other risk scores which necessitate more detailed clinical measurement and venesection for measurement of biochemical and genetic variables, such as the San Antonio Heart Study [5] and Framingham Offspring models [6], have been developed and evaluated. These scores undoubtedly have a greater predictive ability than simple methods such as the Finnish Risk Score when

considered in isolation from response rate but can really only be considered as tools to be used in a clinical, rather than a public health setting. In a clinical context they perform well, particularly when they include glucose, since risk prediction works best when derangements in the measured variable are a manifestation of early disease. These scores certainly have their place, but that place is in the clinical encounter with a specific individual where there is a specific reason for wishing to assess future risk of diabetes. All too often reports have muddled up this purpose with the more public health approach for which questionnaires like the Finnish risk score are intended. When comparing the predictive ability of instruments designed to predict risk of diabetes we should be clear about the context and the purpose. As with all risk prediction tools, that purpose could range from those relevant to public health, such as stratifying populations for targeting interventions without any additional testing or identification of individuals within a population to invite for more definitive risk assessment, through to those with a more clinical intent, such as estimation of risk to provide prognostic information and computation of likely benefit from a preventive intervention. We should consider the evaluation of the tool as it would be used in the real world, otherwise we risk coming to the wrong conclusion about which is the optimal method of risk prediction for that situation. One cannot compare the usefulness of axes and spades as tools without first working out whether you want to cut down a tree or dig a hole.

Duality of interest The authors declare that there is no duality of interest associated with this manuscript.

References

1. Schulze MB, Simmons R, Griffin SG (2011) Risk prediction tools for identifying individuals at risk for type 2 diabetes. *Epidemiologic Reviews* (in press)
2. Alssema M, Vistisen D, Heymans M et al (2011) The DETECT-2 update of the Finnish diabetes risk score for prediction of incident type 2 diabetes. *Diabetologia*. doi:10.1007/s00125-010-1990-7
3. Rahman M, Simmons RK, Harding AH, Wareham NJ, Griffin SJ (2008) A simple risk score identifies individuals at high risk of developing type 2 diabetes: a prospective cohort study. *Fam Pract* 25:191–196
4. Hippisley-Cox J, Coupland C, Robson J, Sheikh A, Brindle P (2009) Predicting risk of type 2 diabetes in England and Wales: prospective derivation and validation of QDScore. *BMJ* 338:b880
5. Stern MP, Fatchi P, Williams K, Haffner SM (2002) Predicting future cardiovascular disease: do we need the oral glucose tolerance test? *Diab Care* 25:1851–1856
6. Meigs JB, Shrader P, Sullivan LM et al (2008) Genotype score in addition to common risk factors for prediction of type 2 diabetes. *N Engl J Med* 359:2208–2219