A nationwide adaptive prediction tool for coronary heart disease prevention

Tim A Holt and Lucila Ohno-Machado

SUMMARY

Standardised electronic recording of cardiovascular risk factor data collected during primary care delivery could be used to create a new strategy, using an adaptive prediction model, for targeting primary prevention interventions at high-risk individuals. In the short term, this should progressively improve data quality and allow risk modification to be monitored at the population level. In the long term, feedback of data on cardiovascular disease development might enable the model to tailor the recommended interventions more appropriately to the needs of the individual and to adapt to future changes in risk patterns. Ultimately, the inclusion of additional cardiovascular risk factors might enable a richer, more realistic picture of cardiovascular risk profiles to be uncovered. This model may have wider uses in both research and practice, and provides a further incentive for the standardisation of record keeping in primary care.

Keywords: *primary prevention; number needed to screen; coronary heart disease; adaptive learning.*

T A Holt, MRCP, FRCGP, Member, Complexity in Primary Care Group and general practitioner, Whitby. L Ohno-Machado, PhD, MD, Associate Professor of Radiology, Health Sciences and Technology, Decision Systems Group, Brigham and Women's Hospital, Harvard Medical School and Massachusetts Institute of Technology, Boston, MA.

Address for correspondence

Dr Tim Holt, Dale End Surgery, Danby, Whitby, North Yorkshire YO21 2JE. E-mail: tholt@ukonline.co.uk

Submitted: 19 October 2002; Editor's response: 16 January 2003; final acceptance: 11 April 2003.

Introduction

THE National Service Framework (NSF) for coronary heart disease[1] recommends that patients with a greater than 30% risk of developing coronary heart disease in the following 10 years should be treated with a similar priority to those with established disease. Identifying such patients, who lack cardiovascular symptoms, presents a challenge for primary care teams, and the NSF stresses the need for a systematic rather than opportunistic service model, using electronic disease registers and standardised Read codes. There have been doubts about the quality and reliability of data collected in primary care ever since the development of large computer databases during the 1980s,[2] but where coding can be standardised, the ability of existing software, such as MIQUEST, to extract the data anonymously from practice databases provides the opportunity for a nationwide data collection system. Such standardisation is becoming an important means through which the goals of the NSF can be achieved.[3] Patterns of cardiovascular risk among the United Kingdom (UK) population may vary in this century, as they did in the last, through changes in demography, lifestyle patterns, social conditions, modification of risk factors, and through genetic exchange with populations not adequately represented in the Framingham study,[4] the data source on which current predictions are based. In this paper we discuss how an adaptive model might facilitate the identification of high-risk patients, progressively improve data quality, and ultimately adapt to the needs of individuals in a situation of changing coronary heart disease risk.

Adaptive predictive models

Adaptive predictive models are capable of 'learning' to classify cases according to patterns. They use existing data and classification 'gold standards' to construct a model that can predict to which class a new case belongs. They include algorithms, such as logistic regression, which was the basis for the Framingham algorithm, as well as more complex models, such as neural networks,[5] support vector machines,[6] and classification and regression trees.[7] Neural networks, which are widely used in industry for pattern recognition and quality control, comprise input and output layers of processing units, between which hidden layers modify the transmission of information by attributing 'weights' to the various patterns of incoming data. Successful pattern recognition is reinforced through an increase in the weight attributed to the relevant input pattern. This is done initially by 'training' the network using an existing database, and then (if necessary) allowing the relative accuracy of future predictions to adjust the weighting mechanism in the hidden layers (Box 1).

A neural network might be used as a quality control device in a plate factory. The inputs would include features of the plate, such as its thickness, reflectivity and shape, and the output would be a prediction of how easily it might break.

The relative success of the predictions (determined by the rate of plate breakage) could be allowed to modify the weights given to the appropriate input patterns, so that the network effectively 'learns' from experience, and can adapt its predictions over time to consistent changes in the environment to which the plates are exposed.

Box 1. An example of a neural network.

- Age
- Sex
- Smoking status
- Systolic blood pressure
- Total serum cholesterol level
- Serum HDL cholesterol level
- Presence or absence of diabetes
- Presence or absence of left ventricular hypertrophy

Box 2. The Framingham input variables.

Baxt has described a neural network model used to interpret patterns of symptoms, clinical signs, and electrocardiograph findings in 356 patients presenting with acute chest pain to a hospital emergency department, 120 of whom were subsequently found to have myocardial infarction.[8] Twenty input variables for each patient were fed in to the network, which was trained using the data from half the patients and then tested on the other half. The training and testing was then repeated using the opposite halves of the sample. The network was able to recognise the patients with myocardial infarction with greater success than either physicians or previous computer-based strategies. By recognising the significance of combinations of minor variables, it performed well even in the absence of electrocardiographic signs of infarction. A neural network model has also been used successfully to assess cardiovascular risk using a number of different lipids as input variables.[9]

If data collected during the process of care is used to build an adaptive prediction tool, it will be more likely that the model will perform well in classifying new cases. Data collected in more controlled environments may be more adequate to characterise risk factors and classify new cases in a similar population than to classify new cases in a different population. The Framingham algorithm has been validated in northern Europeans,[10] but may not remain valid indefinitely, and is not universally applicable to all ethnic groups without recalibration.[11]

One of the strengths of collecting electronic data during the process of care is that large sets of data can be collected in a relatively short time. Taking advantage of this to construct adaptive models that will be used in a similar population is very important. Another benefit is that different models can be constructed in which only certain variables need to have corresponding values. For example, in the electronic implementations of the Framingham algorithm suggested by the NSF,[12] the software will not produce a risk estimate unless all the variables have corresponding values. If the value for HDL cholesterol is unknown, for example, the algorithm either assumes an estimated value or it will not run. Adaptive predictive models built with a subset of the variables could be used in these cases. This ability makes such models useful in the presence of incomplete data, and would become important if additional risk variables were to be included in the calculation in the future.

Current primary prevention strategies

'Ten-year risk' is currently assessed using the Framingham algorithm and the individual patient's risk variables (Box 2). These variables were selected for the Framingham study because they are 'objective and strongly and independently related to CHD'.[13] Other factors known to affect risk include the patient's ethnic group, exercise level, alcohol consumption, other dietary variables, family history, body mass index, and waist-to-hip ratio. The exclusion of these factors limits the accuracy of the Framingham calculation, but in an individual's case can be used to modify risk estimations at the discretion of the clinician.[14] How much to adjust remains an open issue.

The NSF recommends that patients known to have hypertension and/or diabetes are selected first for risk assessment.[1] Such patients are at higher risk than the general population, and the 'number needed to screen' to find an individual with more than a 30% 10-year risk is therefore reduced through this strategy. However, those at highest risk tend to be the older patients in all risk factor groups, and the effect of age may outweigh the other major factors. In the age group 35–39 years, the number needed to screen is greater than 1,000 for both men and women, but only 10 for men and 75 for women aged between 60 and 64 years.[15] Eighty-five per cent of the population's avoidable cardiovascular disease is to be found in the 16% who are over 65 years old.[16] Clinical intuition is not a sufficient means of reducing the number needed to screen, and subjective estimates of individual risk by general practitioners or practice nurses are inferior to computer-assisted risk calculations.[17]

A new targeting strategy

Candidates for primary prevention screening could be identified electronically by roughly estimating the 10-year risk on all patients in the practice, based on the most recent values of the existing coded risk variables, or, in the case of systolic blood pressure, an average of the last two measurements — this is the mechanism used by the current EMIS system to calculate the risk of individual patients. Those patients on treatment for hypertension or hyperlipidaemia would need to be identified with a lower threshold, because they will have a higher risk when assessed using pretreatment levels. Existing computer software in primary care can make such a distinction electronically. While the Framingham algorithm is designed to predict outcomes using pretreatment blood pressure and lipid levels, the same algorithm might be used as a starting point for assessing modified risk and then adjusted according to outcomes using the adaptive prediction model. This would provide essential information on the impact of treatment on risk which is not available from the Framingham study.

In this way, the computer could, through regular searches, identify patients who were actually or potentially drifting into the greater-than-30% range. The practice could then be informed, perhaps on a 3-monthly basis, of all such patients, who would be identified anonymously using electronic record numbers and listed in order of suspected risk. The interval could be adjusted according to available time and resources.

So far, this process could all be carried out at practice level without the need for extraction of data by an external agency, but the pooling of data nationally would have one further potential benefit: adaptive learning of the prediction algorithm. The healthcare system in the UK, as opposed to the United States, is equipped to quickly build predictive models from data collected in the process of care, including models that take into account regional differences in terms of patient population and practice variation.

Adaptive learning

Linking practices by pooling extracted data would, in principle, enable the adaptive prediction model to adjust its internal parameters in response to observed outcomes (namely, the development of coronary heart disease and stroke). This 'reprogramming' is possible because the same database that provides the values of the Framingham variables also contains the dates when each patient who later developed coronary heart disease was diagnosed. The ability of existing computer software to examine data retrospectively on the timing of coronary heart disease onset has already been demonstrated.[18] In principle, therefore, all the information needed to retrain the model is present within the system (Figure 1).

An example of where such modification might occur concerns the predictive values of systolic and diastolic blood pressures, and pulse pressure in relation to age. There is recently published evidence from the Framingham study that diastolic blood pressure is a more reliable predictor of future cardiovascular outcomes in younger patients compared with older ones, in whom systolic pressure is more reliable.[19] Above a certain age, pulse pressure may then become the best predictor. An adaptive predictive model would eventually produce the best prediction it could for each age group when exposed to enough data over extended time periods, recognising that the weights appropriate for the systolic and diastolic blood pressure values would be partly dependent on the value of the age variable.

Advantages of an adaptive prediction tool built with primary care data

The targeting of individuals for risk assessment would be improved by using expected overall risk as the basis for patient selection, rather than a diagnosis of diabetes or hypertension. The electronic retrieval of any of the other variables, the most important of which is age, would assist in reducing the number needed to screen.

Patients who are not diagnosed with hypertension but who have raised blood pressure measurements, and who represent a significant case volume,[20] would be included in the screening process because they would be identified by their blood pressure values, and not on the basis of inclusion in the hypertension disease register.

Where data are missing, a different predictive model could be used (although data should become increasingly complete over time within the higher risk groups).

By measuring risk using the most recent input variable values, the model can monitor the adequacy of risk modification in a practice population, making it amenable to audit. Decisions about treatment can still be based on pretreatment blood pressure and lipid levels, as recommended in the NSF.

The cyclical nature of the process, like the traditional audit cycle, means that improvements are progressive, and patients moving into the high-risk category over time can be recognised. High-risk patients are a dynamic subgroup that is constantly revising its membership. This dynamism needs to be reflected through a targeting policy that is ongoing rather than a 'once-only' exercise.

Those patients at high risk, whose blood pressure defies reduction to target levels through drug treatment can, nevertheless, have their overall risk reduced by the use of combined approaches. This process is facilitated through the monitoring of modified rather than pretreatment risk.

Discussion

The targeting phase of this model has no minimum quality requirement other than an electronic age and sex register, but the adjustment of the algorithm would only be appropriate if data quality were maintained at a high level, creating numerous difficulties. In particular, the measurement of blood pressure would need to be carried out by adequately trained staff, in line with recommended practice.[21] Blood pressure measurements taken by primary care clinicians in busy surgeries, and on patients who may be unwell at the time, may differ from those gathered in the less pressured conditions of a prospective cohort study. Coded outcome measures would need to include all cardiovascular events, including sudden cardiovascular deaths, while morbidity registers for coronary heart disease in general practice are currently of variable quality.[22] Recorded dates of the onset of cardiovascular disease may be delayed following presentation while investigations are undertaken to confirm the diagnosis. Other influences might also undermine the model's validity; for example, financial incentives based on achievement of blood pressure targets rather than on the quality of data recording. Patients moving from one practice to another would need to be identifiable in order to match predictions with outcomes, and might be lost in the process. This problem of 'data censoring'[23] can be accounted for in some of the statistical models proposed above, in order that the information is still useful even if incomplete, but it will remain an issue.

It is therefore likely that some of the participating practices across the UK, with a commitment to maintaining high-quality data and accurate, up-to-date disease registers for both coronary heart disease and diabetes, would need to be identified (Box 3) in order to minimise these obstacles. It might be hoped that the usefulness of the tool would motivate participants to enter high-quality data. The sheer quantity of information available, which would soon exceed any past cohort study, might address questions previously unanswerable owing to inadequate sample sizes. Other

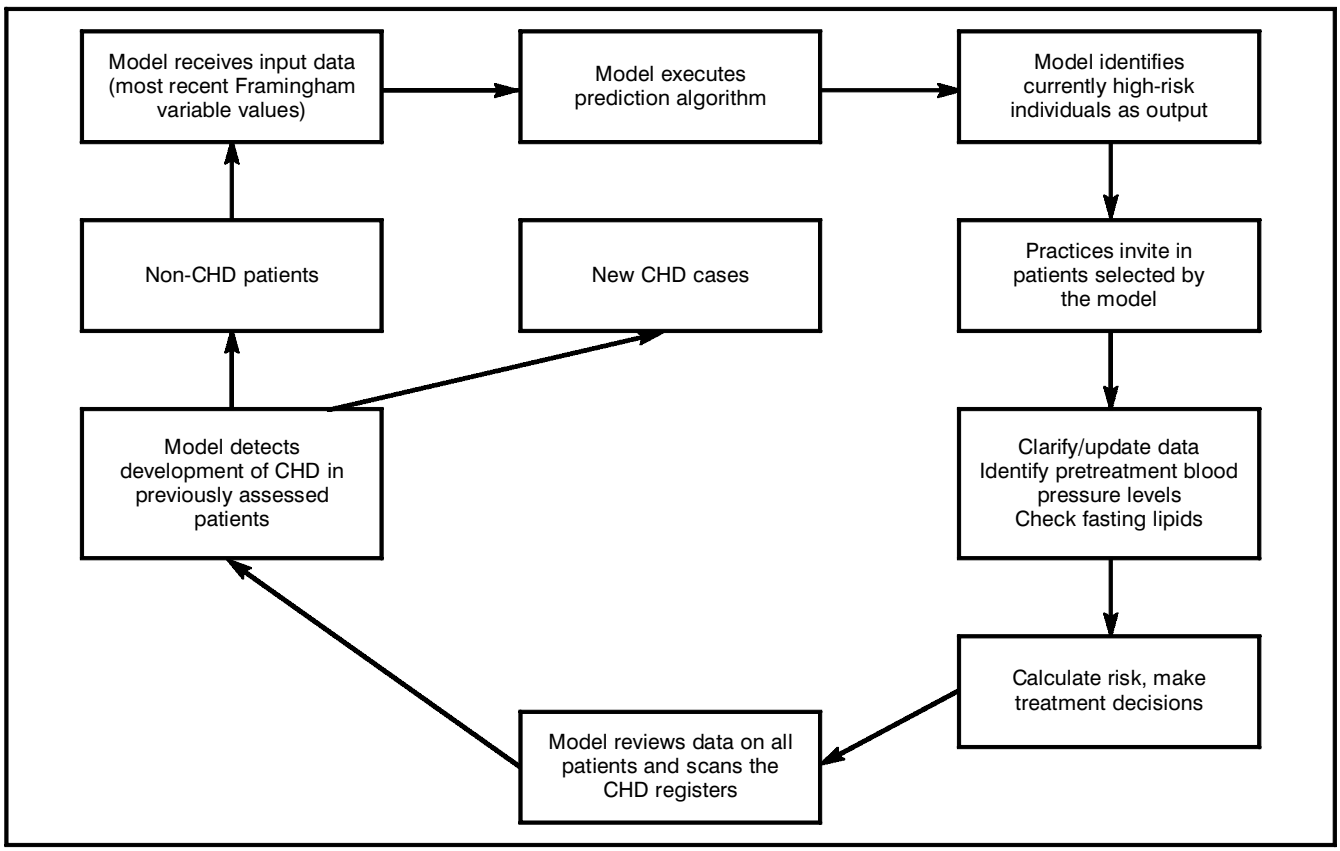

Figure 1. A nationwide adaptive prediction model for coronary heart disease (CHD) prevention. The development of new CHD cases in the population is detected by scanning the CHD register during each cycle. These outcomes are then allowed to modify the prediction algorithm, if necessary, by comparing past risk factor patterns with outcomes in both CHD and non-CHD cases.

- Routine electronic recording of all the Framingham input variables
- Separate electronic disease registers for type 1 and type 2 diabetes
- A coronary heart disease register with accurate dates of onset
- Blood pressure measurements carried out by adequately trained personnel
- A policy of testing for diabetes in patients undergoing cholesterol estimation
- Electronic recording of anti-hypertensive and lipid-lowering medication use
- Coded recording of deaths from cardiovascular disease

Box 3. Minimum data quality standards for participating practices.

factors known to influence risk could be included, and the minor variables might become more important when present in certain combinations, as seen in the example from Baxt discussed above.[8] Such combinations might occur rarely, even in a large cohort study population, and their predictive value may therefore escape recognition. In principle, the model could use any relevant factor that is recordable electronically, including the use of drugs such as aspirin, angiotensin-converting enzyme (ACE) inhibitors, and beta-blockers, as well as the known missing factors discussed above.

The adaptive prediction model would need to be poised to respond to changing patterns with an appropriate sensitivity, in order that only statistically significant trends are allowed to lead to modification of the algorithm. It might be expected that an adjustment in the algorithm would occur initially as a result of risk differences between the original Framingham cohort and the current UK population. Thereafter, more gradual changes might be seen as an adaptation to demographic and genetic changes in the UK population.

The model could only be as 'smart' as the data allowed, and unless given information on ethnicity (which is not routinely recorded electronically), it could not allow for the known differences in risk between different ethnic groups. In practice, however, human involvement, which is of course invaluable to the process of communicating risk and advising on lifestyle modification and treatment to individual patients, would still, under this proposed strategy, allow this and other missing factors to be taken into account when planning treatment, as recommended under the current policy. Coronary heart disease prevention is a challenging area of primary care. This model can only assist in certain stages of a complex process, but might enable resources to be targeted more effectively, advice to become more sensitively tailored to the individual, and in the process generate information for research through a novel mechanism involving practising clinicians in the natural environment of everyday care.

Coronary heart disease prevention is an obvious example where a framework for standardised electronic recording has been specified in the NSF, and a prediction algorithm is already in widespread use. Other potential applications include the assessment of predictive values for primary care symptom complexes[24] and the prognosis of malignant disease in individual patients.[25]

Conclusion

The development of computerised disease registers and the electronic recording of values for cardiovascular risk factor variables open up the possibility of a nationwide adaptive prediction tool, which would be capable of pooling data from a large number of participating practices committed to high-quality data recording. Such a model would function as a pattern recognition device, identifying candidates for coronary heart disease risk assessment and allowing risk control to be monitored at the population level.

In principle, the model could improve the accuracy of predictions currently made through the Framingham algorithm over time, by responding to significant trends in the patterns of coronary heart disease risk in the UK as they develop during the 21st century. Where data quality allows, the same method could be applied to other areas of clinical care, and may help to bridge the gap between research and practice. This provides a further stimulus for the integration and standardisation of electronic record keeping in primary care.

References

1. Department of Health. *National service framework for coronary heart disease.* London: Department of Health, 2000.
2. Pringle M, Hobbs R. Large computer databases in general practice. *BMJ* 1991; **302:** 741-742.
3. Simpson D, Nicholas J, Cooper K. The use of information technology in managing patients with coronary heart disease. *Informatics in Primary Care* 2002; **10:** 15-18.
4. Kannel WB, McGee D, Gordon T. A general cardiovascular risk profile: the Framingham Study. *Am J Cardiol* 1976; **38:** 46-51.
5. Bishop CM. *Neural networks for pattern recognition.* Oxford: Oxford University Press, 1995.
6. Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. In: Haussler D (ed). *Proceedings of the fifth annual ACM workshop on computational learning theory.* New York: ACM Publications, 1992: 144-152.
7. Breiman L, Friedman J, Olshen RA, Stone CJ. *Classification and regression trees.* New York: Chapman & Hall, 1984.
8. Baxt WG. Complexity, chaos and human physiology: the justification for non-linear neural computational analysis. *Cancer Lett* 1994; **77:** 85-93.
9. Lapuerta P, Azen SP, LaBree L. Use of neural networks in predicting the risk of coronary artery disease. *Comput Biomed Res* 1995; **28:** 38-52.
10. Haq IU, Ramsay LE, Yeo WW, *et al.* Is the Framingham risk function valid for northern European populations? A comparison of methods for estimating absolute coronary risk in high risk men. *Heart* 1999; **81(1):** 40-46.
11. D'Agostino RB Sr, Grundy S, Sullivan LM, *et al*. CHD Risk Prediction Group. Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. *JAMA* 2001; **286(2):** 180-187.
12. Hingorani AD, Vallance P. A simple computer program for guiding management of cardiovascular risk factors and prescribing. *BMJ* 1999; **318(7176):** 101-105.
13. Anderson KM, Wilson PWF, Odell PM, Kannel WB. An updated coronary risk profile. A statement for health professionals. *Circulation* 1991; **83(1):** 356-362.
14. Wood D, Durrington P, Poulter N, *et al*. Joint British recommendations on prevention of coronary heart disease in clinical practice. *Heart* 1998; **80(suppl 2):** 1S-29S.
15. Scottish Intercollegiate Guidelines Network. *Lipids and the primary prevention of coronary heart disease.* (Publication 40.) Edinburgh: Scottish Intercollegiate Guidelines Network, 1999.
16. Marshall T, Rouse A. Meeting the National Service Framework for coronary heart disease: which patients have untreated high blood pressure? *Br J Gen Pract* 2001; **51:** 571-574.
17. McManus RJ, Mant J, Meulendijks CF, *et al.* Comparison of estimates and calculations of risk of coronary heart disease by doctors and nurses using different calculation tools in general practice: cross sectional study. *BMJ* 2002; **324(7335):** 459-464.
18. Meal AG, Pringle M, Hammersley V. Time changes in new cases of ischaemic heart disease in general practice. *Fam Pract* 2000; **17(5):** 394-400.
19. Franklin SS, Larson MG, Khan SA, *et al.* Does the relation of blood pressure to coronary heart disease risk change with ageing? The Framingham Heart Study. *Circulation* 2001; **103(9):** 1245-1249.
20. Colhoun HM, Dong W, Poulter NR. Blood pressure screening, management and control in England: results from the health survey for England 1994. *J Hypertens* 1998; **16(6):** 747-752.
21. O'Brien ET, Petrie JC, Littler WA, *et al.* Blood pressure measurement: recommendations of the British Hypertension Society 3rd ed. London: BMJ Publishing Group, 1997.
22. Moher M, Yudkin P, Turner R, *et al*. An assessment of morbidity registers for coronary heart disease in primary care. ASSIST (ASSessment of Implementation STrategy) trial collaborative group. *Br J Gen Pract* 2000; **50:** 706-709.
23. Ohno-Machado L. Modeling medical prognosis: survival analysis techniques. *J Biomed Inform* 2001; **34(6):** 428-439.
24. Summerton N. Symptoms of possible oncological significance: separating the wheat from the chaff. *BMJ* 2002; **325:** 1254-1255.
25. Black N. Using clinical databases in practice. *BMJ* 2003; **326:** 2-3.