PharmacoEconomics (2016) 34:195–205
DOI 10.1007/s40273-015-0342-3

ORIGINAL RESEARCH ARTICLE

Using Classification and Regression Trees (CART) to Identify Prescribing Thresholds for Cardiovascular Disease

Chris Schilling[1] · Duncan Mortimer[2] · Kim Dalziel[1] · Emma Heeley[3] · John Chalmers[3] · Philip Clarke[1]

Published online: 17 November 2015

Abstract

Background and Objective Many guidelines for clinical decisions are hierarchical and nonlinear. Evaluating if these guidelines are used in practice requires methods that can identify such structures and thresholds. Classification and regression trees (CART) were used to analyse prescribing patterns of Australian general practitioners (GPs) for the primary prevention of cardiovascular disease (CVD). Our aim was to identify if GPs use absolute risk (AR) guidelines in favour of individual risk factors to inform their prescribing decisions of lipid-lowering medications.

Methods We employed administrative prescribing information that is linked to patient-level data from a clinical assessment and patient survey (the AusHeart Study), and assessed prescribing of lipid-lowering medications over a 12-month period for patients ($n = 1903$) who were not using such medications prior to recruitment. CART models were developed to explain prescribing practice. Out-of-sample performance was evaluated using receiver operating characteristic (ROC) curves, and optimised via pruning.

Results We found that individual risk factors (low-density lipoprotein, diabetes, triglycerides and a history of CVD), GP-estimated rather than Framingham AR, and sociodemographic factors (household income, education) were the predominant drivers of GP prescribing. However, sociodemographic factors and some individual risk factors (triglycerides and CVD history) only become relevant for patients with a particular profile of other risk factors. The ROC area under the curve was 0.63 (95 % confidence interval [CI] 0.60–0.64).

Conclusions There is little evidence that AR guidelines recommended by the National Heart Foundation and National Vascular Disease Prevention Alliance, or conditional individual risk eligibility guidelines from the Pharmaceutical Benefits Scheme, are adopted in prescribing practice. The hierarchy of conditional relationships between risk factors and socioeconomic factors identified by CART provides new insights into prescribing decisions. Overall, CART is a useful addition to the analyst's toolkit when investigating healthcare decisions.

Key Points for Decision Makers

Classification and regression trees (CART) provide a methodology to highlight how and why variation between practice and guidelines occurs, not just that an evidence-practice gap exists.

Prescribing practices for lipid-lowering medications do not follow absolute risk guidelines or eligibility criteria for subsidisation by the Pharmaceutical Benefits Scheme. There are potentially significant gains from clarifying best-practice prescribing, to promote either greater adherence to guidelines or increased clinical freedom.

Big data techniques such as CART are applicable to a wide range of healthcare applications, including those where big data are absent.

✉ Chris Schilling
chris.schilling@unimelb.edu.au

[1] Centre for Health Policy, School of Population and Global Health, University of Melbourne, Melbourne, VIC 3051, Australia

[2] Centre for Health Economics, Monash Business School, Monash University, Melbourne, VIC 3800, Australia

[3] The George Institute for Global Health, The University of Sydney and the Royal Prince Alfred Hospital, Sydney, NSW 2050, Australia

1 Introduction

Physicians employ a range of risk assessment strategies when making prescribing decisions for primary prevention of cardiovascular disease (CVD), including assessment against thresholds on individual CVD risk factors and assessment of total or absolute risk (AR) of cardiovascular (CV) events [1, 2].

Internationally, many clinical practice guidelines recommend calculation of AR of CV events, with lipid-lowering medication (typically statins or statins in combination with another drug) recommended for patients evaluated as high risk.[1] In Australia, an AR approach is recommended by the National Vascular Disease Prevention Alliance (NVDPA) [5] and the National Heart Foundation (NHF) [6]. However, the Australian Government's universal drug insurance scheme, the Pharmaceutical Benefits Scheme (PBS), limits the subsidising of these medicines using eligibility criteria based on individual risk factors such as diabetes and cholesterol.

In practice, several studies suggest that clinicians deviate from guidelines and/or eligibility criteria [7, 8], perhaps in response to care-seeking behaviour from patients or other aspects of patient preference [1, 2]. However, Bonner et al. [1] noted that while CVD risk management is not consistently based on AR, "little is known about … and the alternative strategies employed when AR is not the focus of assessment". Similarly, few studies have formally tested whether different prescribing thresholds are being applied in different patient groups, and many have struggled to characterise the complexity of prescribing practice [1, 9].

The purpose of this paper was to employ classification and regression trees (CART) to analyse the prescribing patterns of Australian general practitioners (GPs) of lipid-lowering medication for the primary prevention of CVD. CART is a machine-learning 'big data' technique that has been shown to be particularly valuable when analysing nonlinear relationships and interactions, where it can outperform standard regression models for classification [10]. We aimed to use CART to improve our understanding of clinical practice, potentially identifying prescribing thresholds and patient subgroups missed by traditional analyses, and to demonstrate how CART can be useful for understanding complex treatment decisions in healthcare.

[1] For example, the American Heart Association (AHA) recommends using a modified Framingham equation [3]. In the UK, the National Institute for Health and Care Excellence (NICE) recommends an absolute CVD risk algorithm known as QRISK2 [4].

2 Methods

2.1 Data

We used linked survey and administrative data from the AusHeart Study, a cluster-stratified, cross-sectional survey of CVD risk management in primary care fully documented elsewhere [7, 11, 12]. The study enrolled GPs from across Australia, who recruited 15–20 consecutively presenting adults aged 55 years or older. It gathered information on patient socioeconomic factors, CVD risk factors, prescribed medications, and the GP's own estimation of the patient's AR of a CV event within the next 5 years [7].

For consenting patients, these data were linked to Medicare administrative data containing records of all pharmaceuticals purchased under the PBS from 1 March 2008 to 1 January 2010 [12]. To avoid complications associated with prior exposure to medication, we reduced the dataset to 1903 patients who had not been prescribed lipid-lowering medication prior to GP recruitment.[2] We developed models to classify these patients according to prescription/nonprescription of any lipid-lowering medication (Anatomical Therapeutic Chemical code C10) during a 1-year period.

2.2 Classification and Regression Trees (CART) Methodology

CART sorts observations into increasingly homogeneous subgroups [13]. At each step, CART splits observations using a simple decision rule (e.g. if total cholesterol exceeds 7.0 mmol/L, then prescribe medication) chosen to minimise diversity (with respect to the binary outcome or classification) in right and left 'child nodes'. Branches and nodes are added until a stopping criteria is met and the tree terminates in 'leaves' or 'bins' containing proportions of correctly and incorrectly classified observations [10].

There are three distinct strengths of CART that make it particularly applicable to analysing complex decision-making processes such as those employed in clinical practice. First, the hierarchical structure of CART models is often more intuitive than traditional regression models because it mimics the heuristics of decision making [14, 15]. Second, CART can outperform standard regression models when predicting outcomes in the presence of nonlinear relationships and interactions [10]. In clinical practice, treatment decisions may depend on nonlinear thresholds with respect to one or more risk factors, and

[2] Prior exposure to medication is not preferred as we would observe risk factors after response to treatment.

thresholds may vary with other risk factors. For example, PBS guidelines allow prescribing of statins for patients with hypercholesterolemia (>9 mmol/L) [16]. This drops to >5.5 mmol/L if the patient has diabetes [16].[3] Third, CART affords the data greater freedom to speak for themselves [20]. Whereas regression models are refined by comparing across a limited number of possible specifications, CART performs an exhaustive search over all possible cut-points and predictors [10]. As a result, the precise form of the relationship between a predictor and outcome is not delimited by the inclusion/exclusion of higher order terms. It is this strength that has seen CART used in a variety of prognostic analyses to identify risk thresholds for in-hospital mortality [21], vertebral fractures [22] and cirrhosis [23].

However, CART is subject to a number of limitations. CART "… tends not to work very well if the underlying relationship is linear" [10]. A second limitation of CART is the risk of overfitting [24, 25]. Finally, CART can be prone to instability. Small differences in the training data can lead to very different trees [26]. We manage these limitations in the methodology below.

2.3 Using CART to Understand Prescribing in Cardiovascular Disease (CVD)

We used a three-stage approach to construct the CART. In CART-1, we limited predictors to patient sociodemographics (age, sex, Aboriginal/Torres Strait Islander, household income and education level) and GP-estimated 5-year AR of a CV event. This provided a benchmark for which to compare performance. In CART-2, we added individual risk factors (smoking, body mass index, systolic and diastolic blood pressure, low- and high-density lipoprotein cholesterol, total cholesterol, triglycerides, kidney disease, diabetes, CVD history, weekly exercise and self-reported health) "… to determine whether cardiovascular risk factors might have an additional influence on prescribing beyond their contribution to [GP-estimated] cardiovascular risk" [27]. Finally, in CART-3 we added AR, estimated using the 1991 Framingham risk equations. Framingham AR forms the basis of the NHF 2004 and NVDPA 2008 guidelines. If GPs adopt NHF or NVDPA guidelines, we would expect the addition of Framingham AR to improve the predictive validity of the CART, and to see cut-off thresholds and a hierarchy similar to the guidelines (described in "Appendix 1").

We implemented CART using the Matlab *fitctree* function [28]. Gini's diversity index was used as the default splitting criterion, as suggested by Breiman et al. [24], and we compared model performance under entropy splitting to check model robustness. In our default models, variables with missing data still enter the model, but training uses only valid values. In prediction, an observation with a missing value is assigned to the largest split group. An alternative method for dealing with missing data in CART is to find 'surrogate' variables by applying CART with the missing data as the dependent variable [28]. We checked model performance under these two methods to test robustness, and used tenfold cross-validation to indirectly evaluate out-of-sample performance.[4] We bootstraped the cross-validation 100 times to describe the distribution of mean out-of-sample error and receiver operating characteristic (ROC) area under the curve metric. We pruned the CART to reduce overfitting and optimise out-of-sample performance. This helps to eliminate illogical branches that can grow from the sample data but which would not perform well out-of-sample (e.g. where a node suggests that patients with a household income between AUS$52,000 and AUS$72,799 are less likely to be prescribed than those with a household income below AUS$52,000 or above AUS$72,799). Where there was no difference in out-of-sample performance, we followed Breiman et al. [24] in preferring smaller trees over larger trees.

Once optimised, the structure of the CART was evaluated to identify patient subgroups and prescribing thresholds. We calculated a predictor-importance metric for the preferred model using the *predictor-importance* Matlab algorithm.[5] Next, we compared patient subgroups and prescribing thresholds identified by the CART against NHF 2004, NVDPA 2008 and PBS guidelines to identify similarities and differences.

Finally, we evaluated the stability of our results. The robustness of predictor-importance and specific hierarchies is difficult to assess because of the conditional nature of the CART [30]. As a simple guide, we trained 100 'bagged' trees[6] on bootstrapped samples of the data, and counted the number of times each predictor appeared [32, 33]. Following Dannegger [32], we calculated confidence intervals (CIs) and density functions of the cut-off thresholds used at key decision nodes to highlight stability.

[3] Similar complications exist in clinical decision making in general [17], and in observed (as well as recommended) prescribing patterns for statins [1, 18, 19].

[4] This has been shown to be an optimal method for model selection [29].

[5] This identifies all the nodes where the predictor is selected, sums the improvement in classification from each of these and divides by the number of tree branches [28].

[6] Bagging or 'bootstrapped aggregating' is a method for generating multiple versions of a tree to allow evaluation of predictor stability [31].

3 Results

3.1 Prescribing and Risk Factor Statistics

Table 1 provides the sample mean and standard deviation (SD), or frequency count and percentage, for demographic and clinical characteristics. Of the 1903 patients, 296 (16 %) were prescribed lipid-lowering medication.

3.2 Model Performance

CART-1 considers only patient demographics and the GP-estimated 5-year AR of a CV event. It provides a performance benchmark but is not expected to perform well given the absence of individual risk factors or Framingham AR. The unpruned CART-1 correctly identified 1560 (97 %) patients who were not prescribed lipid-lowering medication, but only 115 (39 %) of those who were prescribed, for an overall within-sample error rate of 12 %. As expected, the performance of CART-1 drops when moving out-of-sample; error increases to 20 % (95 % CI 20–21) but with pruning this is reduced to 18 % (95 % CI 17–18). The out-of-sample ROC metric is 0.53 (95 % CI 0.51–0.55), indicating the model is barely better than a random guess at predicting prescribing patterns (Table 2).

CART-2 adds 13 individual risk factors to CART-1. This improves both within- and out-of-sample performance. Within sample, the model correctly identified 1585 (99 %) patients who were not prescribed lipid-lowering medication, and 157 (53 %) of those who were, for an overall error rate of 8 %. After pruning, the out-of-sample error was 17 % (95 % CI 16–17 %) and the ROC metric was 0.63 (95 % CI 0.60–0.64).

CART-3 adds Framingham AR to CART-2, which should identify NHF and/or NVDPA guidelines if they are followed. Within sample, the model correctly identified 1579 (98 %) patients who were not prescribed, and 172 (58 %) who were prescribed, for an overall error rate of 8 %. After pruning, the out-of-sample error was 17 % (95 % CI 16–18) and the ROC metric was 0.62 (95 % CI 0.60–0.63), which is not significantly different from CART-2. Framingham AR does not appear in the pruned version of CART-3.

3.3 Predictors of Prescribing

Household income, GP-estimated AR, and the individual risk factors low-density lipoprotein [LDL], diabetes, total cholesterol, CVD history and triglycerides all influence GP prescribing under the pruned CART-2 model. The predictor-importance results suggest that LDL, GP-estimated AR and diabetes make the most improvement to classification,

Table 1 Characteristics of the patients in the AusHeart study

Variable	Total ($n = 1903$)	CART model
Prescribing		Dependent variable
Lipid-lowering medication	296 (16 %)	
Sociodemographic variables		Explanatory variables in all models
Age (years)	66 ± 9	
Female	1131 (59 %)	
Aboriginal/Torres Strait Islander	16 (1 %)	
Household income (annual)		
Negative/nil	28 (1 %)	
AUS$1–18,199	401 (21 %)	
AUS$18,200–33,799	466 (24 %)	
AUS$33,800–51,999	253 (13 %)	
AUS$52,000–72,799	166 (9 %)	
AUS$72,800–103,999	124 (7 %)	
AUS$104,000 or more	101 (5 %)	
Missing	364 (19 %)	
Education		
None/very little	527 (28 %)	
School/diploma	901 (47 %)	
University	435 (23 %)	
Missing	40 (2 %)	
Individual risk factors		Explanatory variables in models 2 and 3
Current smoker	163 (9 %)	
Body mass index (kg/m^2)	28.2 ± 5.6	
Missing	55 (3 %)	
Systolic blood pressure (mmHg)	136 ± 17	
Diastolic blood pressure (mmHg)	77 ± 10	
Low-density lipoprotein cholesterol (mmol/L)	3.22 ± 0.84	
High-density lipoprotein cholesterol (mmol/L)	1.47 ± 0.45	
Total cholesterol (mmol/L)	5.36 ± 0.93	
Missing	26 (1 %)	
Triglycerides (mmol/L)	1.49 ± 0.82	
Missing	30 (2 %)	
Kidney disease	69 (4 %)	
Diabetes	250 (13 %)	
Missing	3 (0 %)	
CVD history		
None	1618 (85 %)	
Stroke/TIA only	170 (9 %)	
CAD only	86 (5 %)	
Both stroke/TIA and CAD	29 (2 %)	
Exercise per week (>30 min moderate)		
None	352 (18 %)	
1–2 days/week	541 (28 %)	
3–4 days/week	523 (27 %)	
5–7 days/week	439 (23 %)	
Missing	48 (3 %)	

Table 1 continued

Variable	Total ($n = 1903$)	CART model
Self-rated health		
Excellent	124 (7 %)	
Very good	508 (27 %)	
Good	841 (44 %)	
Fair	353 (19 %)	
Poor	47 (2 %)	
Missing	30 (2 %)	
Absolute risk assessments		
GP-estimated absolute 5-year risk (%)	14 ± 17	Explanatory variable in all models
Missing	182 (10 %)	
Framingham absolute 5-year risk (%)	10 ± 7	Explanatory variables in model 3
Missing	87 (5 %)	
Patient self-reported absolute 5-year risk (%)	33 ± 23	

Data are expressed as mean ± SD or frequency counts (%)

CART classification and regression trees, *CVD* cardiovascular disease, *GP* general practitioner, *SD* standard deviation, *TIA* transient ischemic attack, *CAD* coronary artery disease

followed by triglycerides, income, total cholesterol and CVD history (Table 3).

Figure 1 shows interactions between the AR assessments, individual risk factors, and sociodemographic factors, and highlights the paths that lead to prescribing. On the right-hand side of the tree, prescribing is most likely for patients with high LDL (>4.09 mmol/L), high total cholesterol (>6.95 mmol/L), high GP-estimated AR (>17.5 %) and relatively low household income (<AUS$52,000). On the left-hand side of the tree, patients without high LDL (<4.09 mmol/L) are more likely to be prescribed if they have high triglycerides (≥4.25 mmol/L for patients with diabetes; ≥4.35 mmol/L for patients without diabetes and with GP-estimated AR ≥2.5 %), or if they have previously had a coronary artery event.

CART also highlights interactions where prescribing is unlikely. Patients with high LDL, total cholesterol and GP-estimated AR are less likely to be prescribed lipid-lowering medication if they have relatively high household income (≥AUS$52,000). Patients with high LDL and high total cholesterol but without high GP-estimated AR are also less likely to be prescribed. Finally, prescribing is less likely for patients with high LDL but without high total cholesterol.

3.4 Robustness of Results

Comparison of CART-2 performance under different splitting criteria and approaches to missing data show no significant differences in ROC out-of-sample performance (Table 2). Comparison of the 100 bagged trees highlights robustness of the specific hierarchies and decision nodes within CART-2. LDL and diabetes appear in all 100 trees, at the root and second node positions, and have the highest average predictor-importance (Table 3). The LDL decision threshold is bimodal, with a mode at 4.6 mmol/L in addition to the 4.1 mmol/L suggested in CART-2 (Fig. 2); however, the difference between modes is less than one SD in LDL in the sample (0.8 mmol/L). By contrast, total cholesterol appears at the third node in only 6 of the 100

Table 2 Model performance (%)

Model	CART-1		CART-2		CART-3		CART-2 robustness check	
Demographics	All		All		All		All	
Individual risk factors	None		All		All		All	
Absolute risk factors	GP-estimated		GP-estimated		GP-estimated, Framingham		GP-estimated	
Pruning	None	Pruned	None	None	Pruned	Pruned	Pruned	Pruned
Splitting criterion	GDI	GDI	GDI	GDI	GDI	GDI	Entropy	GDI
Missing data	Default	Default	Default	Default	Default	Default	Default	Surrogates
Within-sample error	0.12	0.15	0.08	0.08	0.15	0.15	0.14	0.14
Sensitivity	0.97	0.99	0.99	0.99	1.00	1.00	0.99	1.00
Specificity	0.39	0.09	0.53	0.53	0.07	0.07	0.10	0.09
Out-of-sample error	0.20	0.18	0.22	0.22	0.17	0.17	0.16	0.17
95 % lower bound	0.20	0.17	0.22	0.22	0.16	0.16	0.16	0.16
95 % upper bound	0.21	0.18	0.23	0.23	0.17	0.17	0.17	0.17
Out-of-sample ROC area	0.53	0.53	0.57	0.57	0.63	0.63	0.62	0.61
95 % lower bound	0.50	0.51	0.55	0.55	0.60	0.60	0.60	0.58
95 % upper bound	0.56	0.55	0.59	0.59	0.64	0.64	0.64	0.64

CART classification and regression trees, *GP* general practitioner, *GDI* Gini's diversity index, *ROC* receiver operating characteristic

Table 3 Predictor results

Predictor	Predictor-importance		Counts in bagged trees		Threshold[a]
	Pruned CART-2	Bagged trees	Nodes 1–3	Nodes 4–10	
LDL	1.31	1.59	100	54	4.3 mmol/L (4.1:4.6)
GP-estimated AR	1.05	0.38	0	116	30 % (2.5:80.5)
Diabetes	0.94	0.56	100	0	Yes
Triglycerides	0.47	0.33	1	171	2.4 mmol/L (0.4:4.4)
Income	0.31	0.21	3	73	AUS$52,000
Total cholesterol	0.28	0.08	6	31	5.2 mmol/L (3.6:7.0)
CVD history	0.11	0.08	0	63	Both
Education	0	0.21	44	9	University
Framingham AR	0	0.08	0	10	10.6 % (1.6:30.6)
Exercise per week	0	0.04	0	38	No exercise
Self-rated health	0	0.09	0	4	Very good

CART classification and regression trees, *CVD* cardiovascular disease, *AR* absolute risk, *GP* general practitioner, *LDL* low-density lipoprotein

[a] Confidence intervals for continuous variables; median threshold for discrete variables

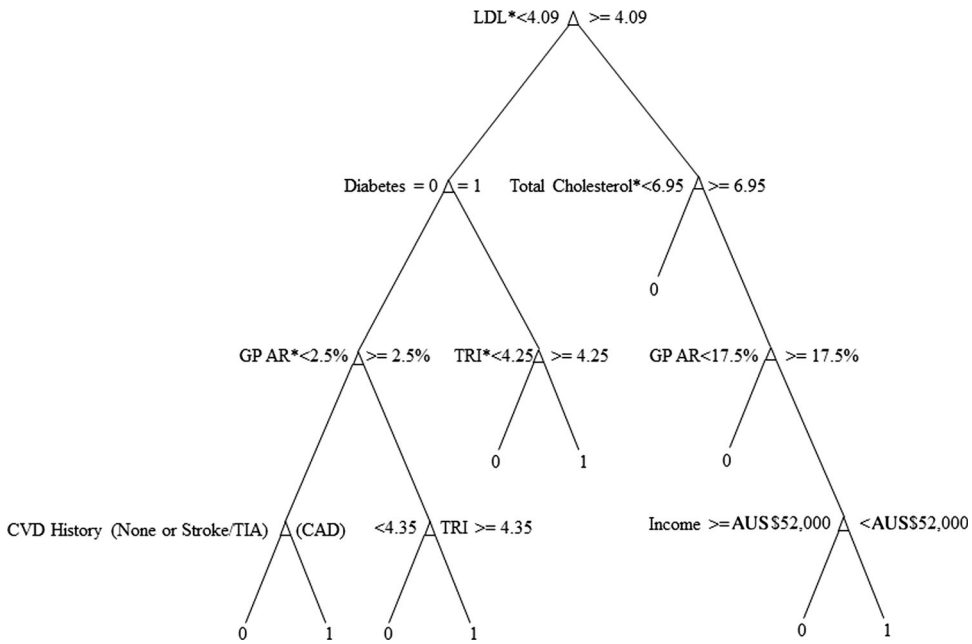

Fig. 1 CART-2. *CART* classification and regression trees, *LDL* low-density lipoprotein (mmol/L), *total cholesterol* total cholesterol (mmol/L), *GP AR* general practitioner-estimated absolute risk (5), *TRI* triglycerides (mmol/L), *CVD* cardiovascular disease, *CAD* coronary artery disease, *TIA* transient ischaemic attack, *Income* household income, *Asterisk* 0 indicates not prescribed medication, 1 indicates prescribed medication

bagged trees. Education (those with University education are less likely to be prescribed) appears 44 times at node 3. Triglycerides and GP-estimated AR, which appear twice in CART-2, appear 171 and 116 times, respectively, within the first 10 nodes. The triglyceride decision threshold shows the cut-off at 4.3 mmol/L, as seen in CART-2, but also identifies another mode at 2.0 mmol/L. Household income appears 76 times in the first 10 nodes, with the median cut-off at AUS$52,000 as per CART-2. Exercise, Framingham AR and self-rated health status are not present in the pruned CART-2 model but appear in 38, 10 and 4 of the first 10 nodes of the 100 bagged trees.

4 Discussion

4.1 Key Findings

4.1.1 Prescribing Varies Across GPs and Does Not Appear to Follow AR Guidelines or PBS Regulations

We find that prescribing practices do not appear to be congruent with NHF, NVDPA or PBS eligibility guidelines. NHF and NVDPA use Framingham AR assessment as the basis of their guidelines, yet thresholds on Framingham AR rarely appear in the CART. The guidelines also

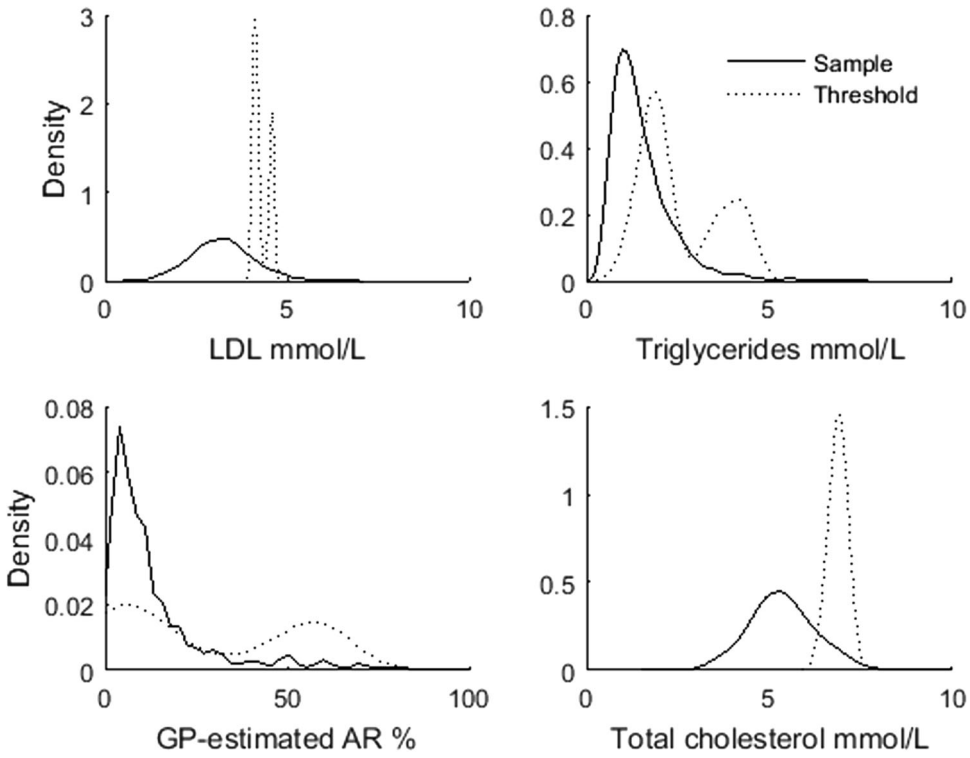

Fig. 2 Sample and threshold densities for selected risk factors. *AR* absolute risk, *GP* general practitioner, *LDL* low density lipoprotein

recommend prescribing on the basis of individual risk factors (e.g. for patients with kidney disease or diabetes). Kidney disease does not appear in CART-2; however, diabetes does appear but is neither necessary nor sufficient for prescription.

Similarly, the PBS has conditional criteria based on individual risk that govern eligibility, such as total cholesterol >5.5 mmol/L for patients with diabetes, and >6.5 mmol/L for patients with HDL <1 mmol/L. These decision branches do not appear in CART-2; however, the model suggests that prescribing is more likely for low LDL patients with triglycerides >4.25 mmol/L (for patients with diabetes) and >4.35 mmol/L (for patients with high GP-estimated AR). This is somewhat consistent with the PBS, which allows prescribing for a subset of patients with triglycerides >4 mmol/L.

Our findings contribute to a growing body of evidence [2, 7, 18, 27], suggesting there is considerable room for improvement in the prescribing practices for CVD. If guidelines provide an accurate description of optimal treatment, divergence from guidelines is likely to be costly, both in terms of health expenditure and patient outcomes. For example, the prescription of drugs to patients who fall outside the specified indications, often referred to as leakage [34], is likely to diminish the real-world cost effectiveness of pharmaceuticals if it results in patients gaining a lower average benefit than was assumed at the time of the approval for use. There may then be dividends from interventions to improve adherence to guidelines, such as IMPLEMENT, ALIGN and IRIS [35–37]. CART would be an appropriate method to assess such adherence. Conversely, if thresholds for reimbursement constrain best-practice prescribing (e.g. based on total or absolute risk of CV events or a more thorough understanding of the patient), there may be a case for removing thresholds for reimbursement and allowing increased clinical freedom in prescribing. Either way, there are potentially significant opportunity costs to this uncertainty.

While we find discordance between practice and guidelines, we do not identify one standard decision tree that consistently explains prescribing behaviour across our representative dataset. Instead, we find that prescribing practices vary across the GP population. This is perhaps unsurprising given the volume of guidance available [38] and the potential for between-GP variation in uptake and acceptance of decision-support tools and guideline recommendations [39]. We posit that the low ROC performance of the CART models is a result of this variation. In an environment of clearer and more widely adopted guidelines, we would expect the ROC performance to improve.

4.1.2 CART Suggests How and Why GP Prescribing Deviates From Guidelines

The CART analysis provides additional insights regarding the roles of individual risk factors and the hierarchy of GP decision making. LDL is the root node in all bagged trees, suggesting it is the first risk factor used in the decision-making process. Similarly, diabetes is consistently the second node in the decision tree, suggesting it is an important risk factor that GPs consider in decision making. It is well established that lowering CV risk is associated with the degree to which statins reduce LDL cholesterol [40]. Similarly, statins have been widely prescribed in people with diabetes, given their higher CV risk [41]. It appears that evidence regarding these risk factors takes precedence over AR and the eligibility criteria.

We also show that prescribing to high-risk patients varies based on the patient's household income and/or educational attainment, with those with household income above AUS$52,000 or with a university degree unlikely to be prescribed. There has been some evidence of this internationally [42, 43]; however, the CART method uncovers the hierarchy of these factors. Specifically, we show that income/educational attainment are deciding factors at the bottom of the prescribing decision tree, after clinical establishment of high risk. However, there is likely to be confounding between these factors and patient's health and lifestyle. Self-rated health and exercise are significantly collinear with household income (Pearson Chi-square p values of 0.016 and 0.000), and both entered some trees in the robustness analysis. Nonetheless, the results concord with theories that GPs use clinical judgement and knowledge of the patient to make decisions based on a wide range of factors, not just AR-based guidelines involving absolute or individual risk factors [1].

Finally, we show that CVD history is taken into consideration for otherwise low-risk patients, with those patients with CAD more likely to be prescribed. This concurs with previous research that highlighted inconsistent CV risk perceptions across vascular territories [44].

4.2 Limitations

There are limitations to this study. First, the analysis uses filled prescriptions, rather than written prescriptions, as the measure of prescribing. To the extent that patients with unfilled scripts differ in some respect from more compliant patients, the CART may not characterise prescribing practice across all patient groups.[7] Caution should therefore be exercised in generalising our findings to patients prescribed but who do not go on to fill their prescriptions.

[7] For example, there is some evidence to suggest that compliance increases with the number of risk factors [45].

Second, the AusHeart sample is a stratified random sample of GPs who had previously expressed an interest in participating in the study. While this approach produced a nationally representative sample with respect to a number of observable GP characteristics [7], it may have selected GPs with a greater than average interest in CVD management and the guidelines associated with it. There are also some limitations from the survey design; for example, we do not know the time interval of prescribing or nonprescribing prior to the study.

Finally, the CART method has limitations. Overall model performance is low, which could be due to variance in prescribing practices; each GP might use a different tree for each patient. We discuss GP variability and clustering in "Appendix 2". Instability in trees uncovered by CART can be difficult to measure and visualise.

5 Conclusions

While previous studies showed discordance between evidence and practice, CART extends traditional analyses by highlighting the alternative decision trees and key factors that GPs use in practice to make prescribing decisions. The advantages of CART are the ability to identify hierarchies and nonlinearities, and to provide results that are relatively easy to understand. These strengths are evident in this analysis, which show hierarchical decisions with complex interactions between individual risk factors and sociodemographic factors.

This example has shown that the CART big data technique is applicable to a wide range of healthcare topics, including those where big data are absent. There are an increasing range of applications in healthcare that utilise CART's strength in uncovering nonlinear thresholds and hierarchies to develop guidelines for clinical decisions. It follows that evaluating if these guidelines are used in practice requires methods that can identify such structures and thresholds. In these instances, CART provides a useful addition to the analyst's toolkit.

Acknowledgments This work was supported by Monash University, the George Institute for Global Health, and the University of Melbourne.

Chris Schilling, Duncan Mortimer, Kim Dalziel, Emma Heeley, John Chalmers and Philip Clarke declare that they have no conflicts of interest.

Author contributions CS, DM and KD conceptualized this report. All authors had input in developing the approach. CS produced multiple drafts. All authors provided input on the draft report and all read and approved the final report.

Appendix 1

See Table 4.

Table 4 Summary of Australian guidelines current during the study period, to inform the prescribing of lipid-lowering medication

Guideline	Guidance to inform prescribing of lipid-lowering medication	Use of absolute risk?
NHF, 2008[a]	Use of lifestyle modification first-line	Yes, Framingham 1991 equation
	Use of drugs for:	
	Those with existing disease (vascular, diabetes, kidney or hypercholesterolemia), Aboriginal Torres Strait Islander people, and those with AR ≥15 % in 5 years	
	Those with AR of 10–15 % and a family history of CHD, or with metabolic syndrome	
NVDPA, 2009	Adults with any of the following high-risk conditions:	Yes, use of Framingham (for patients not assessed as high risk using other criteria)
	Diabetes and age >60 years	
	Diabetes with microalbuminuria (>20 μg/min or urinary albumin:creatinine ratio >2.5 mg/mmol for males or >3.5 mg/mmol for females)	
	Moderate or severe CKD (persistent proteinuria or eGFR <45 mL/min/1.73 m^2)	
	A previous diagnosis of familial hypercholesterolaemia[a]	
	Systolic blood pressure ≥180 mmHg or diastolic blood pressure ≥110 mmHg	
	Serum total cholesterol >7.5 mmol/L	
	Use of Framingham to assess risk in those not considered 'high risk'	
PBS, 2014	Dietary therapy should be trialled prior to drug therapy for all patients who are not very high risk	No, evidence included the Heart Protection Study (HPS), the United Kingdom Prospective Diabetes Study (UKPDS), Australian data audits and input from experts
	Use of drugs for:	
	Very-high-risk patients (e.g. existing CHD, vascular disease or diabetes, or a family history of CHD)	
	Patients not considered very high risk who have combinations of risk (e.g. diabetes, Aboriginal, high HDL cholesterol, family history) along with particular lipid levels, e.g. an Aboriginal patient with total cholesterol >6.5 mmol/L. See PBS (2014) for complete details	

NHF National Heart Foundation, *NVDPA* National Vascular Disease Prevention Alliance, *PBS* Pharmaceutical Benefits Scheme, *AR* absolute risk, *CHD* coronary heart disease, *CKD* chronic kidney disease, *eGFR* estimated glomerular filtration rate, *HDL* high-density lipoprotein

[a] Guideline refers readers to the National Heart Foundation of Australia and the Cardiac Society of Australia and New Zealand Position Statement on Lipid Management (2005)

Appendix 2

After condensing the data to obtain a single observation per patient, our CART makes no further adjustment for clustering of observations by GP. On average, GPs see eight patients within the dataset (minimum of one patient per GP; maximum of 16). Stability across bagged trees may be overestimated if 'bags' of observations are drawn from clustered data. In supplementary analyses, we evaluated stability of the CART in 100 samples drawn using cluster-bootstrap methods [46]. Predictor counts and threshold densities were much the same with the cluster bootstrap as for the simple bootstrap on clustered data described above.

Similarly, while detailed contextual data on each GP was not available, the data did contain a State location variable that identifies the GP's geographic region. In supplementary analyses, we included this variable within the predictor set, however it did not enter into the preferred CART model shown in Fig. 1.

References

1. Bonner C, et al. General practitioners' use of different cardiovascular risk assessment strategies: a qualitative study. Med J Aust. 2013;199(7):485–9.
2. Jansen J, et al. General practitioners' use of absolute risk versus individual risk factors in cardiovascular disease prevention: an experimental study. BMJ Open. 2014;4(5):e004812.
3. Greenland P, et al. 2010 ACCF/AHA guideline for assessment of cardiovascular risk in asymptomatic adults: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines developed in collaboration with the American Society of Echocardiography, American Society of Nuclear Cardiology, Society of Atherosclerosis Imaging and Prevention, Society for

Cardiovascular Angiography and Interventions, Society of Cardiovascular Computed Tomography, and Society for Cardiovascular Magnetic Resonance. J Am Coll Cardiol. 2010;56(25):e50–103.
4. Stone NJ, et al. 2013 ACC/AHA guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. J Am Coll Cardiol. 2014;63(25 Pt B):2889–934.
5. Lalor E, et al. National Vascular Disease Prevention Alliance. Guidelines for the management of absolute cardiovascular disease risk. 2012. ISBN 978–0–9872830–1–6. https://strokefoundation.com.au/~/media/strokewebsite/resources/treatment/absolutecvd_gl_webready.ashx?la=en.
6. National Heart Foundation of Australia. Guide to management of hypertension 2008. 2010. Available at: http://www.heartfoundation.org.au/SiteCollectionDocuments/HypertensionGuidelines2008to2010Update.pdf.
7. Heeley EL, et al. Cardiovascular risk perception and evidence-practice gaps in Australian general practice (the AusHEART study). Med J Aust. 2010;192(5):254–9.
8. Razavian M, et al. Cardiovascular risk management in chronic kidney disease in general practice (the AusHEART study). Nephrol Dial Transpl. 2012;27(4):1396–402.
9. Zwar N, et al. GPs' views of absolute cardiovascular risk and its role in primary prevention. Aust Fam Physician. 2005;34(6):503–4.
10. Varian HR. Big data: new tricks for econometrics. J Econ Perspect. 2014;28(2):3–27.
11. Knott RJ, et al. How fair is Medicare? The income-related distribution of Medicare benefits with special focus on chronic care items. Med J Aust. 2012;197:625–30.
12. Knott RJ, et al. The effects of reduced copayments on discontinuation and adherence failure to statin medication in Australia. Health Policy. 2015;119(5):620–7.
13. Hothorn T, et al. Party: a laboratory for recursive partytioning. 2010. Available at: https://cran.r-project.org/web/packages/party/vignettes/party.pdf.
14. Drakopoulos SA. Hierarchical choice in economics. J Econ Surv. 1994;8(2):133–53.
15. Scott A. Identifying and analysing dominant preferences in discrete choice experiments: an application in health care. J Econ Psychol. 2002;23(3):383–98.
16. Pharmaceutical Benefits Scheme. General statement for lipid-lowering drugs prescribed as pharmaceutical benefits. Pharmaceutical Benefits Scheme; 2014.
17. Garg AX, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. JAMA. 2005;293(10):1223–38.
18. Berthold H, et al. Patterns and predictors of statin prescription in patients with type 2 diabetes. Cardiovasc Diabetol. 2009;8(1):25.
19. Wong M, et al. Patterns of antihypertensive prescribing, discontinuation and switching among a Hong Kong Chinese population from over one million prescriptions. J Hum Hypertens. 2008;22(10):714–6.
20. Timofeev R. Classification and regression trees (CART) theory and applications. Humboldt-Universitat zu Berlin, Wirtschaftswissenschaftliche Fakultat. 2004. http://edoc.hu-berlin.de/docviews/abstract.php?id=26951.
21. Tomcikova D, et al. Epidemiology, quality improvement and outcome: risk of in-hospital mortality identified according to the typology of patients with acute heart failure: classification tree analysis on data from the Acute Heart Failure Database–main registry. J Crit Care. 2013;28:250–8.
22. Navarro Mdel C, et al. Discriminative ability of heel quantitative ultrasound in postmenopausal women with prevalent vertebral fractures: application of optimal threshold cutoff values using classification and regression tree models. Calcif Tissue Int. 2012;91(2):114–20.
23. Shi K-Q, et al. Risk stratification of spontaneous bacterial peritonitis in cirrhosis with ascites based on classification and regression tree analysis. Mol Biol Rep. 2012;39(5):6161–9.
24. Breiman L, et al. Classification and regression trees. Boca Raton: CRC Press; 1984.
25. Torgo L. Inductive learning of tree-based regression models. Universidada do Porto. Reitoria. 1999. http://repositorio-aberto.up.pt/handle/10216/10018.
26. Briand B, et al. A similarity measure to assess the stability of classification trees. Comput Stat Data Anal. 2009;53(4):1208–17.
27. Mohammed MA, El Sayed C, Marshall T. Patient and other factors influencing the prescribing of cardiovascular prevention therapy in the general practice setting with and without nurse assessment. Med Decis Making. 2012;32(3):498–506.
28. The Mathworks Inc. Matlab and statistics toolbox release 2015a. Natick: The Mathworks; 2015.
29. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. IJCAI. 1995;2:1137–43.
30. Kitsantas P, Hollander M, Li LM. Assessing the stability of classification trees using Florida birth data. J Stat Plan Inference. 2007;137(12):3917–29.
31. Breiman L. Bagging predictors. Mach Learn. 1996;24(2):123–40.
32. Dannegger F. Tree stability diagnostics and some remedies for instability. Stat Med. 2000;19(4):475–91.
33. Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychol Methods. 2009;14(4):323.
34. Lopert R, Henry D. The Pharmaceutical Benefits Scheme: economic evaluation works… but is not a panacea. Aust Prescr. 2002;25(6):126.
35. French SD, et al. Evaluation of a theory-informed implementation intervention for the management of acute low back pain in general medical practice: the IMPLEMENT cluster randomised trial. PLoS One. 2013;8(6):e65471.
36. McKenzie JE, et al. Evidence-based care of older people with suspected cognitive impairment in general practice: protocol for the IRIS cluster randomised trial. Implement Sci. 2013;8(1):91.
37. McKenzie JE, et al. Improving the care for people with acute low-back pain by allied health professionals (the ALIGN trial): a cluster randomised trial protocol. Implement Sci. 2010;5(1):86.
38. Allen D, Harkins KJ. Too much guidance? Lancet. 2005;365(9473):1768.
39. Prevedello LM, et al. Does clinical decision support reduce unwarranted variation in yield of CT pulmonary angiogram? Am J Med. 2013;126(11):975–81.
40. Cholesterol Treatment Trialists' (CTT) Collaborators. Efficacy and safety of cholesterol-lowering treatment: prospective meta-analysis of data from 90,056 participants in 14 randomised trials of statins. Lancet. 2005;366(9493):1267–78.
41. Kearney P, et al. Efficacy of cholesterol-lowering therapy in 18,686 people with diabetes in 14 randomised trials of statins: a meta-analysis. Lancet. 2008;371(9607):117–25.
42. Ashworth M, et al. Social deprivation and statin prescribing: a cross-sectional analysis using data from the new UK general practitioner 'Quality and Outcomes Framework'. J Public Health (Oxf). 2007;29(1):40–7.
43. Weitoft GR, et al. Education and drug use in Sweden—a nationwide register-based study. Pharmacoepidemiol Drug Saf. 2008;17(10):1020–8.

44. Heeley E, et al. Disparities between prescribing of secondary prevention therapies for stroke and coronary artery disease in general practice. Int J Stroke. 2012;7(8):649–54.
45. Latry P, et al. Adherence with statins in a real-life setting is better when associated cardiovascular risk factors increase: a cohort study. BMC Cardiovasc Disord. 2011;11(1):46.
46. Field CA, Welsh AH. Bootstrapping clustered data. J R Stat Soc Series B Stat Methodol. 2007;69(3):369–90.