Journal of Evaluation in Clinical Practice
International Journal of Public Health Policy and Health Services Research

Journal of Evaluation in Clinical Practice ISSN 1365-2753

Using data mining techniques to characterize participation in observational studies

Ariel Linden DrPH[1,2] and Paul R. Yarnold PhD[3]

[1]President, Linden Consulting Group, LLC, Ann Arbor, MI, USA
[2]Research Scientist, Division of General Medicine, Medical School, University of Michigan, Ann Arbor, MI, USA
[3]President, Optimal Data Analysis, LLC, Chicago, IL, USA

Keywords
data mining, machine learning, observational studies, observed characteristics, selection bias, selection

Correspondence
Dr Ariel Linden
Linden Consulting Group, LLC
1301 North Bay Drive
Ann Arbor, MI 48103
USA
E-mail: alinden@lindenconsulting.org

Accepted for publication: 4 January 2016

doi:10.1111/jep.12515

Abstract

Data mining techniques are gaining in popularity among health researchers for an array of purposes, such as improving diagnostic accuracy, identifying high-risk patients and extracting concepts from unstructured data. In this paper, we describe how these techniques can be applied to another area in the health research domain: identifying characteristics of individuals who do and do not choose to participate in observational studies. In contrast to randomized studies where individuals have no control over their treatment assignment, participants in observational studies self-select into the treatment arm and therefore have the potential to differ in their characteristics from those who elect not to participate. These differences may explain part, or all, of the difference in the observed outcome, making it crucial to assess whether there is differential participation based on observed characteristics. As compared to traditional approaches to this assessment, data mining offers a more precise understanding of these differences. To describe and illustrate the application of data mining in this domain, we use data from a primary care-based medical home pilot programme and compare the performance of commonly used classification approaches – logistic regression, support vector machines, random forests and classification tree analysis (CTA) – in correctly classifying participants and non-participants. We find that CTA is substantially more accurate than the other models. Moreover, unlike the other models, CTA offers transparency in its computational approach, ease of interpretation via the decision rules produced and provides statistical results familiar to health researchers. Beyond their application to research, data mining techniques could help administrators to identify new candidates for participation who may most benefit from the intervention.

Introduction

With the rapidly growing size and availability of medical data, health researchers are increasingly using data mining tools to handle complex analytical problems [1,2]. Classification is the most popular data mining application in health care and has been used to improve diagnostic accuracy, identify high-risk patients and extract concepts in unstructured data [3].

There is a fundamental difference between the data mining tools used for classification (also referred to as predictive modelling) and conventional statistical modelling methods (e.g. multivariate regression). Specifically, data mining algorithms find the best fitting model through automated processes (called machine learning) that search through the dataset to detect patterns. These patterns may include interactions between variables, as well as interactions within subsets of variables. In conventional statistics, a model is chosen based on an *a priori* hypothesis about the data, and then statistical tests are performed after estimation to verify that the data fit the model [4]. Investigators using conventional statistical methods must manually enter variables, interactions and polynomials, and there is no guarantee that the best fitting model will be discovered.

However, rather than viewing data mining and statistics as rival approaches to classification problems, health researchers may be best served by considering the synergies between them. For example, data mining algorithms can be applied first to identify influential variables and interactions in the data, with the results reviewed and, if needed, refined by a domain expert. Conventional statistics would then be used to assess the model's predictive performance (e.g. area under the receiver operating characteristic curve, effect strength for sensitivity (ESS)] [5,6] on the complete dataset, as well as on hold-out samples in order to test generalizability of the model (e.g. via leave-one-out cross validation, bootstrapping) [7,8]. Measuring predictive performance allows the health researcher to identify the most accurate model among competing approaches, whereas assessing generalizability will help determine how well the model can identify new participants that may not have the identical characteristics as those in the original sample.

In this paper, we describe another area of health research where there are synergies between data mining techniques and conventional statistics: characterizing the individuals who participate in observational studies. In contrast to randomized studies where individuals have no control over their treatment assignment, participants in observational studies self-select into the treatment arm and are therefore likely to differ in their characteristics from those who elect not to participate. These differences may explain part, or all, of the difference in the observed outcome (i.e. selection bias) [9]. Data mining techniques can identify patterns in the data that distinguish study participants from non-participants, revealing potentially complex relationships among individual characteristics that may bias the outcome analysis. From an administrative perspective, the results of data mining as applied to study participation could help identify new candidates for enrolment who may most benefit from the intervention.

To develop and illustrate this approach, the paper is organized as follows: In the second section, we briefly describe the data we use to illustrate how data mining techniques can assist health researchers in identifying the selection issues intrinsic to observational studies. In the third section, we describe two approaches from conventional statistics most commonly used for characterizing selection and assessing potential for bias, as well as describe their limitations. In the fourth section, we describe data mining applications that may be considered for the same purposes. In the fifth section, we discuss approaches to assessing model accuracy in order to determine whether the selected model (from either conventional statistics or data mining) fits the data, and then in the sixth section, we apply these approaches to compare the alternative methodologies (conventional statistics as well as different data mining models). Finally, in the last section, we discuss the implications and applications of our findings.

Data

For exposition, we use data from a primary care-based medical home pilot programme that invited patients to enrol if they had a chronic illness or were predicted to have high costs in the following year. The goal of the programme was to lower health care costs for programme participants by providing intensified primary care (see the study of Linden [10] for a more comprehensive description). The retrospectively collected data consist of observations for 374 programme participants and 1628 non-participants. Eleven pre-intervention characteristics were available; these included demographic variables (age and gender), health services utilization (primary care visits, other outpatient visits, laboratory tests, radiology tests, prescriptions filled, hospitalizations, emergency department visits and home-health visits) and total medical costs (the amount paid for all these health services).

Commonly used approaches for identifying selection

Table of baseline characteristics

The most common approach for identifying selection in intervention studies is by presenting a table of the summary statistics of pre-intervention characteristics for the treatment and control groups [11]. In a sufficiently large randomized trial, it is expected that most, if not all, of the baseline characteristics will be comparable between groups. However, in non-randomized studies where individuals select to participate, there is no expectation that, prior to adjustment, the treatment and control groups will be comparable in their characteristics. Thus, in non-randomized studies, the baseline characteristics table serves to identify the characteristics on which the two groups differ due to selection.

Table 1 presents the observed pre-intervention characteristics of the participants and non-participants in the pilot study [10]. Continuous variables are summarized by mean and standard deviation, and categorical variables are presented as number and percent. For

Table 1 Baseline (12 months) characteristics of programme participants and non-participants [10]

	Participants (*n* = 374)	Non-participants (*n* = 1628)	Standardized differences	*P*-value*
Demographic characteristics				
Age	54.9 (6.71)	43.4 (11.99)	1.704	<0.001
Female	211 (56.4%)	807 (49.6%)	0.138	0.017
Utilization and cost				
Primary care visits	11.3 (7.30)	4.6 (4.35)	0.914	<0.001
Other outpatient visits	18.0 (16.65)	7.2 (10.61)	0.647	<0.001
Laboratory tests	6.1 (5.27)	2.4 (3.31)	0.705	<0.001
Radiology tests	3.2 (4.46)	1.3 (2.48)	0.424	<0.001
Prescriptions filled	40.6 (29.96)	11.9 (17.14)	0.956	<0.001
Hospitalizations	0.2 (0.52)	0.1 (0.29)	0.326	<0.001
Emergency department visits	0.4 (1.03)	0.2 (0.50)	0.226	<0.001
Home-health visits	0.1 (0.88)	0.0 (0.38)	0.083	0.012
Total costs	8236 (9830)	3047 (5817)	0.528	<0.001

*A two-tailed *t*-test for independent samples was used for continuous variables, and a chi-squared test was used for dichotomous variables. Continuous variables are reported as mean (standard deviation) and dichotomous variables are reported as *N* (percent).

Table 2 Logistic regression for predicting participation in the pilot programme [10]

Variable	Odds ratio	Std. Err.	z	P > z	95% CI
Age	1.113	0.011	10.760	<0.001	1.091–1.135
Female	1.091	0.163	0.580	0.560	0.813–1.463
Primary care visits	1.092	0.019	4.950	<0.001	1.054–1.130
Other outpatient visits	1.019	0.007	2.950	0.003	1.006–1.032
Laboratory tests	1.010	0.020	0.490	0.626	0.970–1.050
Radiology tests	1.002	0.021	0.090	0.931	0.960–1.044
Prescriptions filled	1.026	0.003	7.780	<0.001	1.019–1.032
Hospitalizations	1.955	0.515	2.550	0.011	1.167–3.275
Emergency department visits	1.018	0.109	0.170	0.864	0.826–1.255
Home-health visits	0.797	0.078	–2.310	0.021	0.657–0.966
Total costs	1.000	0.000	–0.490	0.627	0.999–1.000

balance measures, we report the standardized difference, for which perfect balance is zero [12], and the conventional *P*-value, where variables with values ≤ 0.05 may be considered imbalanced. It is clear that the participant group differed markedly from the non-participant group on every characteristic. On average, participants were older, were less likely to be female, and had higher utilization and costs than non-participants. All standardized differences were far greater than zero, and all *P*-values were ≤ 0.05.

There are three limitations to this descriptive approach for identifying selection. First, when reviewing summary statistics of individual variables alone, any imbalances in interactions between two or more variables go undetected. Also, although it is possible to generate and test these additional terms, it is both tedious (to account for all pairwise interactions presently, an additional 55 variables would need to be added) and the results may not be very informative if the relationship between the variables is non-linear. Second, this approach provides no way to assess the relative importance of each variable in its contribution to the selection process. For example, in Table 1, we see that there is a very large difference in the mean ages of the two groups (54.9 versus 43.4 for participants and non-participants, respectively). However, there is no information as to whether age is a significantly more important criterion for selecting to participate in the intervention than any other variable. Third, these summary statistics do not indicate whether there is a particular cut-off on a variable's continuum, above which individuals are more likely to participate, or below which individuals are more likely not to participate. This information could be useful when trying to understand what may have caused the intervention to be particularly attractive to certain types of participants.

Logistic regression

Whereas a table of baseline characteristics is descriptive, logistic regression offers a predictive approach to assess selection in observational studies. In such a model, the outcome is the treatment assignment (a binary value of 1 = treatment, and 0 = non-treatment) and it is regressed on all observed baseline characteristics. Thus, this model serves to determine which variables predict treatment status. Table 2 presents the odds ratio of each variable in predicting participation in the pilot. As seen, 6 of the 11 variables significantly ($P < 0.05$) predict participation.

The limitations of this approach are similar to those described in Table 1. First, this model includes only main effects, and thus, interaction terms must be manually generated and tested. Second, there is no simple way of determining the relative importance of each variable in its contribution to predicting participation. Lastly, there is no determination of a cut-off above or below which an individual is more likely to participate.

Data mining tools for identifying selection

Data mining classification tools overcome the limitations of the two common approaches described above, while offering the ability to both *predict* and *explain* selection in observational studies. Additionally, a standard feature of the data mining process involves cross-validating the model to assess its generalizability. This is important if the goal of the analysis is to assist programme administrators to identify new candidates for participation in an ongoing intervention. Cross-validation is less important if the goal is only to estimate treatment effects of the intervention.

There are hundreds of different machine learning algorithms belonging to a large array of classifier families, including discriminant analysis, Bayesian, neural networks, support vector machines (SVMs), decision trees, rule-based classifiers, boosting, bagging, stacking, random forests (RFs) and other ensembles, generalized linear models, nearest neighbours, partial least squares and principal component regression, logistic and multinomial regression, multiple adaptive regression splines and other methods. In an extensive comparison across 121 datasets, Fernández-Delgado *et al*. [13] identified two classifiers out of this large collection that consistently outperformed all other algorithms – SVMs [14] and RFs [15]. We therefore include SVM and RF models in the current paper to represent the best of most commonly used algorithms. Additionally, we include classification tree analysis (CTA) – a classifier that was not examined by Fernández-Delgado *et al*. [13], but which has been shown to consistently outperform other classifiers [8]. As compared to SVM and RF, CTA also provides straightforward interpretable formulae and visual displays to characterize the complex relationship between the covariates and the outcome, enhancing the practical value of data mining methods by elucidating the characteristics of individuals who select to participate in observational studies.

SVMs

The SVM is an algorithm that attempts to find the greatest separation between the two outcome categories (in this case, participation or non-participation in the pilot) on any one or more variables. The line that separates the participants from the non-participants is called the *separating hyperplane*, and the *maximum-margin hyperplane* is that particular hyperplane that defines the greatest separation between the two outcomes. The individual observations closest to the *maximum-margin hyperplane* (on either side of the hyperplane) are called *support vectors*. Thus, the objective is to identify the set of support vectors that uniquely defines the maximum-margin hyperplane for the given data problem [16]. SVM can also provide a solution for non-linearly separable data by applying a kernel function to the data [14].

RFs

To better understand how an RF works, it is first important to describe how an individual *decision tree* – which serves as the basis for the RF – is generated. As the name implies, a decision tree [17,18] is a tree-like structure in which the first variable (root node) is split into branches based on one or more cut-points on that variable, and those branches either point to other variables (child nodes) or terminate at the outcome variable (leaf) when the data cannot be split any further. An individual's predicted classification can be ascertained by simply following the related decision rules from the root node to the terminating leaf.

Constructing a decision tree is an iterative process that involves identifying variables (or chains of variables) that are best able to discriminate between individuals who fall into the different outcome categories. There are several approaches for determining the structure of the decision tree. A widely used decision tree algorithm called C4.5 [18] uses a measure called the *information gain* as a guide. In general, this measure represents the informational value of generating a new branch off of an existing node. The information gain is calculated for all variables, and the node is split on the variable that provides the largest information gain. This procedure continues recursively until either the data cannot be split any further, or until the information gain is zero [17].

An RF is an 'ensemble' method whereby multiple decision trees are grown by a randomized tree-building algorithm [15]. The algorithm involves two bootstrap procedures: individual observations are randomly sampled with replacement and a randomly drawn subset of variables is considered at each node. Thus, RFs achieve superior accuracy over other data mining methods by combining the results of multiple decision trees – each generated under various random conditions.

CTA

CTA is a 'decision-tree'–like classification model that provides accurate, parsimonious decision rules that are easy to interpret (and visually display), while reporting P values derived via permutation tests performed at each node. Two generations of CTA models have been developed for which software is commercially available: hierarchically optimal (HO-CTA) [19,20] and enumerated-optimal (EO-CTA) [8,21]. All CTA models consist of nodes, each representing a variable (also called an *attribute)* selected on the basis of the predictive accuracy it achieves. For each potential variable, a predictive model is identified that maximizes the ESS statistic (described in the fifth section). For an ordered or continuous variable, the model has the form: if score ≤ (value) it predicts that the observation is from outcome class A; otherwise, it predicts that the observation is from outcome class B. For a categorical variable, the model has the form: if score = (category list), it predicts that the observation is from outcome class A; otherwise, it predicts outcome class B. Statistical significance of ESS is evaluated using a permutation probability (no distributional assumptions are made), and a sequentially rejective Sidak–Bonferroni-type multiple comparisons methodology is used to ensure the desired experiment-wise Type I error rate, adjusting for the number of variables (nodes) in the CTA model [8,20]. For CTA, an *a priori* minimum strata N (CTA models terminate in two or more endpoints representing different sample strata) is typically set at 10% of the overall study sample (assuming this provides adequate statistical power) to inhibit over-fitting and increase the likelihood of the model cross-generalizing to independent random samples having comparable or smaller sample size [8].

In HO-CTA, the root node is the variable that yields maximum ESS for the total sample. If the root node achieves ESS = 0 or if $P > 0.05$, then no model can be identified. However, if ESS < 100 and sufficient statistical power exists for sample strata identified by the root node (i.e. A and B in the example above), then the variable yielding the highest ESS with Bonferroni-corrected $P < 0.05$ for each strata is added to the model. Model growth is terminated when accuracy cannot be improved for any branch of the model. After the model is grown, it is pruned in order to explicitly maximize ESS [8,20]. In EO-CTA, the first three nodes are enumerated (all variables satisfying the Bonferroni criterion are enumerated for the first three nodes, and the CTA model yielding maximum *ESS* is retained as the EO-CTA solution).

HO-CTA and EO-CTA models have been developed for many applications and have consistently achieved greater accuracy in training and validity analysis than competing models that maximize variance or the value of the likelihood function [8]. Studies of EO-CTA and HO-CTA applied to the same data showed that EO-CTA models are usually more accurate and more parsimonious – as model accuracy increases, the number of misclassified observations (and statistical power) decreases [8,22].

Assessing classification accuracy

To determine which classification approach is the most accurate, competing models are typically compared across several measures of accuracy. The starting point in assessing a model's accuracy is by presenting the relevant counts of correctly and incorrectly classified observations in a standard classification table (also called a *confusion matrix*), such as that presented in Table 3, and then calculating the following measures: *Sensitivity* (true positive rate) is the proportion of actual participants that are correctly predicted by the model as being participants: D/(C + D) × 100% (Table 3). *Specificity* (true negative rate) is the proportion of actual non-participants that are correctly predicted by the model as being non-participants: A/(A + B) × 100%. The *positive predictive value* (PPV) is the probability that individuals predicted by the model to be participants are true participants: D/(B + D) × 100%. Furthermore, the *negative predictive value* (NPV) is the probability that individuals predicted by the model to be non-participants are true

Table 3 Classification table used for assessing model accuracy (modified from Linden [5])

	Model prediction		
Actual status	Non-participation	Participation	Total
Non-participation	True negative A	False positive B	A + B
Participation	False negative C	True positive D	C + D
Totals	A + C	B + D	A + B + C + D

non-participants: A/(A + C) × 100%. The *overall predictive accuracy* is the proportion of the sample correctly classified: [(A + D)/(A + B + C + D) × 100%].

A perfect classification model would have 100% sensitivity and 100% specificity, thereby correctly identifying all true cases (e.g. participants) and never mislabelling non-cases (e.g. non-participants). In reality, however, few models are that accurate. There are at least three limitations to using these traditional measures of classification accuracy. First, when the test is based on a continuous variable, the sensitivity and specificity will change as the cut-point is moved up or down the continuum. For example, if high values indicate participation, raising the cut-point will mean fewer participants will be correctly classified, thereby decreasing sensitivity but also decreasing the number of false-positives. Conversely, lowering the cut-point results in more participants correctly classified as participants (increased sensitivity), but at the expense of a higher false-positive rate. The second limitation is that the predictive values of the classification model are highly sensitive to the prevalence rate of the observed outcome in the population being evaluated [23]. When the population has a high prevalence of the outcome, the PPV will increase and NPV will decrease. Conversely, when there is low outcome prevalence, PPV decreases and NPV increases. Thus, in a population where nearly everyone is participating in the intervention, it would be much easier to predict a person's likelihood of being in the intervention, and much harder to predict who will be a non-participant. The third limitation (related to the first two limitations) is that these metrics do not provide a consistent directional result. Sensitivity (specificity) may be high while specificity (sensitivity) is low, and PPV (NPV) may be high while NPV (PPV) is low. As a consequence, the researcher may find it difficult to determine which of the measures provide the most meaningful information about the model's accuracy.

Receiving operating characteristic (ROC) curves offer a more robust alternative to these more conventional classification methods. ROC analysis involves first obtaining the sensitivity and specificity of every individual in the sample and then plotting sensitivity versus 1 – specificity across the full range of values. So that one does not have to rely on visual inspection to determine how well the model performs, it is possible to assess the overall classification accuracy by calculating the area under the curve (AUC), also referred to as the *C* statistic. A model with perfect discriminatory ability will have a *C* statistic of 1.0, whereas a model unable to distinguish between participants and non-participants will have a *C* statistic of 0.50. Other levels of discrimination have been proposed. For example, Yourman *et al*. [24] considered *C* statistics in the range of 0.50–0.59 to indicate poor, 0.60–0.69 to indicate moderate, 0.70–0.79 to indicate good, 0.80–0.89 to indicate very good, and 0.90 or greater to indicate excellent discrimination.

The advantages of ROC analysis over conventional 2 × 2 tables are threefold: (1) a pre-determined cut-off point is not required because each possible decision threshold is calculated and incorporated into the analysis; (2) ROC analysis allows for visual examination of scores on one curve or a comparison of two or more curves using the same metric; and (3) the prevalence of the outcome in the sample population is not a limiting factor as it is with the conventional measures of accuracy.

Another measure of accuracy that is superior to the conventional indices described above is the ESS, introduced by Yarnold and Soltysik [6]. ESS is a chance-corrected (0 = the level of accuracy expected by chance) and maximum-corrected (100 = perfect, errorless prediction) index of predictive accuracy. The formula for computing ESS for binary case classification is:

$$\begin{aligned}\text{ESS} = {} & [(\text{Mean percent accuracy in classification} - 50)]/\\ & 50 \times 100\%,\end{aligned} \tag{1}$$

where

$$\begin{aligned}&\text{Mean percent accuracy in classification}\\ &\quad = (\text{Sensitivity} + \text{Specificity})/2 \times 100.\end{aligned} \tag{2}$$

Yarnold and Soltysik [6] considered ESS values less than 25% to indicate a relatively weak, 25%–50% to indicate a moderate, 50%–75% to indicate a relatively strong, and 75% or greater to indicate a strong effect. Using ESS, an investigator may directly compare the performance of different models, relative to chance, regardless of the structural features of the analyses, such as sample size, number of outcome categories, number of covariates and covariate metrics, and sample skew [25]. An advantage that ESS holds over the *C* statistic is that it is easily calculated from values in the classification table [26].

Comparison of approaches for assessing selection in the medical home pilot

Accuracy

In this section, we compare the accuracy between one conventional statistical approach (logistic regression), and the three data mining approaches described earlier, for assessing selection in the medical home pilot data. The 'base case' is the full logistic regression (LR) model presented in Table 2. We then examine SVM and RF, and enumerated CTA. We assess accuracy by comparing: (1) sensitivity, specificity, NPV and PPV; (2) AUC; and (3) ESS, as described previously.

Generalizability

After assessing accuracy, assessing the generalizability of a model helps determine how well it can identify new participants who may have somewhat different characteristics than those in the original sample. Such an application may be valuable for administrators of ongoing interventions who want to limit their enrolment to those individuals who may benefit most from the intervention.

To assess generalizability, a model is first estimated using the entire sample (training set), and accuracy measures are calculated, as described previously. Next, the same model is subjected to a procedure called *cross-validation* and then the accuracy measures are recalculated. If the accuracy measures remain consistent with those of the original model using the entire sample, then we can say that the model is generalizable. Although there are several cross-validation techniques available, the present study utilizes one of two approaches: leave-one-out (LOO) cross-validation is simply n-fold cross-validation, where n is the number of observations in the dataset. Each observation in turn is left out, and the model is estimated for all remaining observations. The predicted value is then calculated for the one hold-out observation, and the accuracy is determined as success or failure in predicting the outcome for that observation. The results of all n predictions are used to calculate the final accuracy estimates displayed in the classification tables, which are then compared to the original estimates [16]. The k-fold cross-validation method is a variant on the LOO approach. Here, the entire dataset is randomly divided into k subsets, the specified model is fit using the other $k-1$ subsets and the resulting parameters are used to predict the outcome in the remaining hold-out subset. Model accuracy measures are calculated using the average values across all hold-out models [16]. As above, the accuracy measures of the cross-validated model are compared to those of the original model using the entire dataset. The model is considered generalizable if the accuracy measures remain consistent with those of the original model.

Model estimation

For all models, the treatment assignment indicator (a binary value of 1 = treatment, and 0 = non-treatment) was the outcome (class), and the 13 observed baseline characteristics were covariates (attributes). All models were estimated using cross-validation. We estimated the LR model using a user-written command for Stata, LOOCLASS [27], which performs LOO and produces the classification measures described earlier. SVM was implemented using the LibSVM library in Weka version 3.7.12 (http://www.cs.waikato.ac.nz/ml/weka/) with a Gaussian kernel and 10-fold cross-validation. The RF algorithm was also implemented in Weka, using a forest of random tree-based classifiers with 100 trees and 10-fold cross-validation. Enumerated CTA was implemented using Automated CTA Software [25]. All variables included in the CTA model were constrained to achieve identical classification accuracy in training (total sample) and LOO validity analysis. To ensure adequate statistical power, inhibit over-fitting, and increase the likelihood of cross-validation if the model is applied to classify a (smaller) independent sample, model endpoints were constrained to have $N \geq 10\%$ of the total sample [8].

Results

Appendices A–D provide the classification tables and accuracy measures for each of the four cross-validated models. As seen, the models performed quite differently across the various measures. For example, LR, SVM and RF had very high specificity (ranging from 93.5% to 95.5%) and low sensitivity (ranging from 37.4% to 52.1% – values reflecting accuracy equal to or less than expected by chance), while CTA had high sensitivity (86.6%) and somewhat lower specificity (75.6%). Additionally, LR, SVM and RF all had very similar NPV values (ranging from 86.9% to 89.5%), whereas CTA had a much higher NPV (96.1%). Conversely, although LR, SVM and RF all had very similar PPV (about 65%), CTA produced a lower PPV (44.9).

It is on the two 'holistic' measures of accuracy – AUC and ESS – where a clear differentiation between models is evident. Here, the CTA achieved a high AUC of 0.81, which categorized it as a 'very good' classifier, whereas the other three models were categorized as only 'moderate' to 'good'. Similarly, CTA achieved a high ESS of 62.3%, categorizing it as 'relatively strong', whereas the other three models were categorized as only 'moderate'.

What is noteworthy about CTA outperforming the other models is that only three covariates were needed to achieve maximal accuracy (office visits, age and prescriptions filled), whereas the other models used all 13 covariates, and achieved lower accuracy. Figure 1 illustrates how those covariates are used in the CTA model to predict participation and non-participation. The model predicted that 47.12% of individuals having more than 7.5 office visits in the previous year were likely to participate in the medical home pilot. Additionally, 38.5% of individuals who had 7.5 or fewer office visits, older than 50.5 years and filled more than 16.5 prescriptions in the previous year were likely to participate. Following the diagram down the left branches, one can see that the model predicted non-participation nearly perfectly with no more than three variables. And finally, from a statistical perspective, health researchers can feel confident in the reliability of the model's discriminatory ability, given that the permutation tests, performed at all nodes, had P values < 0.0001. To summarize, CTA achieved higher accuracy than the competing models, while using much fewer variables. Additionally, all nodes achieved high levels of statistical significance, indicating that the overall model met conventional conditions for demonstrating discriminatory ability as well.

Discussion

Although data mining has broad application in health care, this paper has focused on its use for characterizing the nature of individuals who participate in observational studies. Moreover, as we have demonstrated using the medical home pilot data, CTA holds several advantages over other classification algorithms that may preclude their utility or acceptance by health researchers. First, CTA models offer transparency in the computational approach, unlike the more computationally intensive techniques that offer no interpretable formulae or visual displays of the final model. Second, CTA generally produces parsimonious models that achieve classification accuracy as well as – if not better – than more complex algorithms [8]. As a general rule, a simpler model is always preferred over a more complex model, assuming both

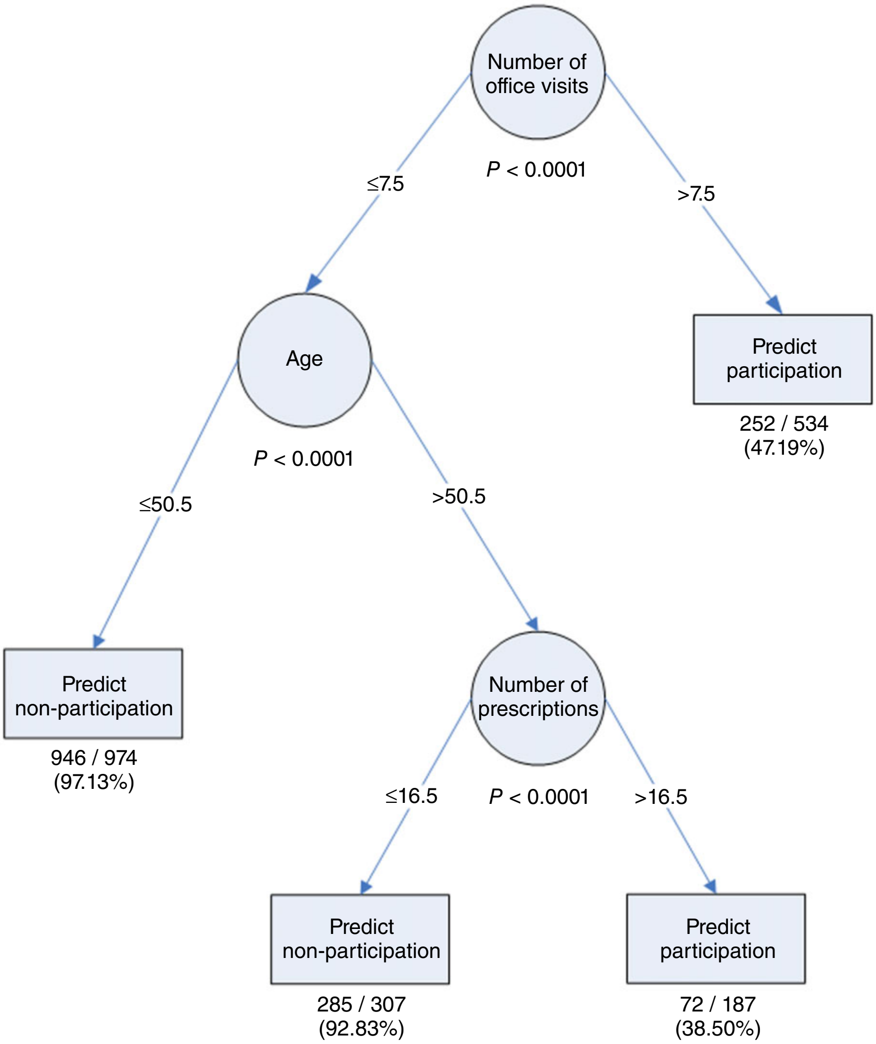

Figure 1 Enumerated classification tree analysis for predicting participation and non-participation in the pilot programme.

have the same classification accuracy. And third, CTA includes permutation tests, adjusted for multiple comparisons, to ensure that the final model meets rigorous statistical assumptions [8,20,21]. Thus, one may consider CTA as an 'all-in-one' classification algorithm that combines the synergies of data mining and conventional statistics. That is, the data mining component ensures that the final model achieves maximum accuracy (as measured by ESS), and the permutation tests, performed at each node, ensure that the model's discriminatory ability has met accepted levels of statistical significance.

The synergies between data mining and statistics can be realized using other models outside of CTA, although it would require substantial manual processing. For example, a decision tree model can be constructed using data mining software, after which the prediction estimates at each node would be retrieved (i.e. the number of correctly and incorrectly predicted cases), and permutation tests on these values would be estimated in a statistical software program. Unfortunately, statistical tests cannot be easily combined with more complex algorithms, such as SVM and RF, which do not provide straightforward interpretable formulae or decision rules. However, it is possible that in the future these models will begin to include statistical testing as a component of their procedures.

For health researchers interested in leveraging data mining to assess selection, the approach should be shaped by the investigator's purpose for determining selection. For example, in the case where an administrator would like to create a tailored recruitment plan targeting individuals who are most likely to benefit from the intervention, a CTA model should be constructed to achieve the highest generalizability while using the fewest variables (using enumerated-optimal CTA). In other words, the model should be accurate for classifying potentially new participants on the basis of a short list of characteristics to minimize the administrative burden. On the other hand, if the investigator wants to reveal potentially complex relationships among individual characteristics that may bias an outcomes study, then a CTA model should be constructed to achieve maximum accuracy without consideration of cross-validation. Additionally, the P values estimated at each node will provide the investigator confidence that the results meet the *a priori* statistical rigour they are willing to accept.

One can also envision how the results from CTA using observational data may inform the inclusion criteria in a prospective trial. Rather than recruiting subjects based on arbitrary cut-off levels on certain covariates (e.g. age over 65), cut-offs can be based on levels most likely to be associated with participation (or non-participation). Subjects may therefore be stratified according to their likelihood of participation, mitigating any subjectivity from the model-selection decision-making process.

The CTA model can also help to identify pathways to further understanding of the phenomenon under investigation, and improve predictive accuracy, vis-à-vis inspection of model residuals (i.e. misclassified observations). For example, in Fig. 1, the two left-hand endpoints yield high predictive accuracy, correctly classifying 97.13% and 92.83% of the observations. In contrast, the two right-hand endpoints yield substantially lower predictive accuracy. Of these latter two endpoints, the right-most misclassifies $n = 282$ observations, and the second-from-the-right endpoint misclassifies $n = 115$ observations. The most efficient follow-up study would therefore obtain and study a sample of observations having eight or more office visits. Classification accuracy for this cohort cannot be improved by using any of the predictive variables (attributes) used presently: a new set of potential predictors is needed. Less efficient but still substantial improvement could be realized by studying a cohort of patients having seven or fewer office visits, age greater than 50.5 years, and 16 or more prescriptions [8].

Although this paper has focused solely on the application of data mining techniques to classifying selection into an observational intervention or treatment, a logical extension of these methods is in the evaluation of outcomes. Athey and Imbens [28] offered a novel conceptual approach for estimating heterogeneous causal effects using data mining techniques. However, this area of research is open to much further exploration. In particular, emphasis should be placed on determining the most appropriate algorithm – or a generalization to all algorithms, extension to outcomes with censored data [29], and the development of specific sensitivity analyses for these applications [30].

In summary, this paper introduced data mining techniques as a novel approach for characterizing the nature of individuals who participate in observational studies. In our motivating example, we found that enumerated CTA achieved the greatest classification accuracy among the models tested, in addition to providing a statistically robust, parsimonious, transparent model.

Acknowledgement

We wish to thank Julia Adler-Milstein for reviewing the manuscript and providing many helpful comments.

References

1. Hand, D. J. (2000) Mining medical data. *Statistical Methods in Medical Research*, 9, 305–307.
2. Smyth, P. (2000) Data mining: data analysis on a grand scale. *Statistical Methods in Medical Research*, 9, 309–327.
3. Iavindrasana, J., Cohen, G., Depeursinge, A., Müller, H., Meyer, R. & Geissbuhler, A. (2009) Clinical data mining: a review. In *IMIA Yearbook of Medical Informatics*, Geissbuhler, A., Kulikowski, C. (editors), 48, Suppl 1, 121–133.
4. Breiman, L. (2001) Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16, 199–231.
5. Linden, A. (2006) Measuring diagnostic and predictive accuracy in disease management: an introduction to receiver operating characteristic (ROC) analysis. *Journal of Evaluation in Clinical Practice*, 12, 132–139.
6. Yarnold, P. R. & Soltysik, R. C. (2005) Optimal Data Analysis: A Guidebook with Software for Windows. Washington, DC: APA Books.
7. Linden, A., Adams, J. & Roberts, N. (2004) The generalizability of disease management program results: getting from here to there. *Managed Care Interface*, 17, 38–45.
8. Yarnold, P. R. & Soltysik, R. C. (2016) Maximizing Predictive Accuracy. Chicago, IL: ODA Books. doi: 10.13140/RG.2.1.1368.3286.
9. Linden, A., Adams, J. & Roberts, N. (2003) An assessment of the total population approach for evaluating disease management program effectiveness. *Disease Management*, 6, 93–102.
10. Linden, A. (2011) Identifying spin in health management evaluations. *Journal of Evaluation in Clinical Practice*, 17, 1223–1230.
11. Linden, A. & Roberts, N. (2005) A users guide to the disease management literature: recommendations for reporting and assessing program outcomes. *American Journal of Managed Care*, 11, 81–90.
12. Linden, A. & Samuels, S. J. (2013) Using balance statistics to determine the optimal number of controls in matching studies. *Journal of Evaluation in Clinical Practice*, 19, 968–975.
13. Fernández-Delgado, M., Cernadas, E., Barro, S. & Amorim, D. (2014) Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, 15, 3133–3181.
14. Noble, W. S. (2006) What is a support vector machine? *Nature Biotechnology*, 24, 1565–1567.
15. Breiman, L. (2001) Random forests. *Machine Learning*, 45, 5–32.
16. Witten, I. H., Frank, E. & Hall, M. A. (2011) Data Mining: Practical Machine Learning Tools and Techniques, 3rd edn. San Francisco, CA: Morgan Kaufmann.
17. Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984) Classification and Regression Trees. Belmont, CA: Wadsworth International Group.
18. Quinlan, J. R. (1993) C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann.
19. Yarnold, P. R., Soltysik, R. C. & Bennett, C. L. (1997) Predicting in-hospital mortality of patients with AIDS-related *Pneumocystis carinii* pneumonia: an example of hierarchically optimal classification tree analysis. *Statistics in Medicine*, 16, 1451–1463.
20. Yarnold, P. R. & Bryant, F. B. (2015) Obtaining a hierarchically optimal CTA model via UniODA software. *Optimal Data Analysis*, 4, 36–53.
21. Yarnold, P. R. & Bryant, F. B. (2015) Obtaining an enumerated CTA model via Automated CTA Software. *Optimal Data Analysis*, 4, 54–61.
22. Feinglass, J., Yarnold, P. R., Martin, G. J. & McCarthy, W. J. (1998) A classification tree analysis of selection for discretionary treatment. *Medical Care*, 36, 740–747.
23. Altman, D. G. & Bland, M. (1994) Diagnostic tests 2: predictive values. *British Medical Journal*, 309, 102.
24. Yourman, L. C., Lee, S. J., Schonberg, M. A., Widera, E. W. & Smith, A. K. (2012) Prognostic indices for older adults: a systematic review. *JAMA: The Journal of the American Medical Association*, 307, 182–192.
25. Soltysik, R. C. & Yarnold, P. R. (2010) Automated CTA software: fundamental concepts and control commands. *Optimal Data Analysis*, 1, 144–160.
26. Linden, A. (2015) CLASSTABI: Stata module for generating classification statistics and table using summarized data. Statistical Software Components s458127, Boston College Department of Economics. Available at: https://ideas.repec.org/c/boc/bocode/s458127.html (last accessed 30 December 2015).
27. Linden, A. (2015) LOOCLASS: Stata module for generating classification statistics of Leave-One-Out cross-validation for binary outcomes. Statistical Software Components s458032, Boston College

Department of Economics. Available at: http://ideas.repec.org/c/boc/bocode/s458032.html (last accessed 23 November 2015).

28. Athey, S. & Imbens, G. (2015) Recursive Partitioning for Heterogeneous Causal Effects. *Working Paper*. Available at: http://arxiv.org/abs/1504.01132 (last accessed 20 January 2016).
29. Linden, A., Adams, J. & Roberts, N. (2004) Evaluating disease management program effectiveness: an introduction to survival analysis. *Disease Management*, 7, 180–190.
30. Linden, A., Adams, J. & Roberts, N. (2006) Strengthening the case for disease management effectiveness: unhiding the hidden bias. *Journal of Evaluation in Clinical Practice*, 12, 140–147.

Appendix A

Classification table and measures of accuracy for a logistic regression model using all 11 baseline variables to predict participation (and non-participation) in the pilot intervention

	Model prediction		
Actual participation	Non-participation	Participation	Total
Non-participation	1541	87	1628
Participation	215	159	374
Total	1751	251	2002

Measure of accuracy	Value (%)
Overall accuracy	84.92%
Sensitivity (participants)	42.51%
Specificity (non-participants)	94.66%
Positive predictive value (participants)	64.63%
Negative predictive value (non-participants)	87.76%
Effect strength (Sensitivity)	37.17%
ROC area	0.69

Appendix B

Classification table and measures of accuracy for a support vector machine (SVM) model to predict participation (and non-participation) in the pilot intervention

	Model prediction		
Actual participation	Non-participation	Participation	Total
Non-participation	1554	74	1628
Participation	234	140	374
Total	1788	214	2002

Measure of accuracy	Value (%)
Overall accuracy	84.62%
Sensitivity (participants)	37.43%
Specificity (non-participants)	95.45%
Positive predictive value (participants)	65.42%
Negative predictive value (non-participants)	86.91%
Effect strength (Sensitivity)	32.89%
ROC area	0.66

Appendix C

Classification table and measures of accuracy for a random forest model to predict participation (and non-participation) in the pilot intervention

	Model prediction		
Actual participation	Non-participation	Participation	Total
Non-participation	1522	106	1628
Participation	179	195	374
Total	1701	301	2002

Measure of accuracy	Value (%)
Overall accuracy	85.76%
Sensitivity (participants)	52.14%
Specificity (non-participants)	93.49%
Positive predictive value (participants)	64.78%
Negative predictive value (non-participants)	89.48%
Effect strength (Sensitivity)	45.63%
ROC area	0.73

Appendix D

Classification table and measures of accuracy for an enumerated classification tree model to predict participation (and non-participation) in the pilot intervention

	Model prediction		
Actual participation	Non-participation	Participation	Total
Non-participation	1231	397	1628
Participation	50	324	374
Total	1281	721	2002

Measure of accuracy	Value (%)
Overall accuracy	77.67%
Sensitivity (participants)	86.63%
Specificity (non-participants)	75.61%
Positive predictive value (participants)	44.94%
Negative predictive value (non-participants)	96.10%
Effect strength (Sensitivity)	62.25%
ROC area	0.81