Journal of Clinical Epidemiology 60 (2007) 1015−1021

Journal of Clinical Epidemiology

The *K*-nearest neighbor algorithm predicted rehabilitation potential better than current Clinical Assessment Protocol

Mu Zhu[a], Wenhong Chen[a], John P. Hirdes[b,c], Paul Stolee[b,d,*]

[a]*Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada*
[b]*Department of Health Studies and Gerontology, University of Waterloo, Waterloo, Ontario, Canada*
[c]*Homewood Research Institute, Homewood Health Centre, Guelph, Ontario, Canada*
[d]*School of Optometry, University of Waterloo, 200 University Avenue W., Waterloo, Ontario, Canada*

Accepted 25 September 2006

Abstract

Objective: There may be great potential for using computer-modeling techniques and machine-learning algorithms in clinical decision making, if these can be shown to produce results superior to clinical protocols currently in use. We aim to explore the potential to use an automatic, data-driven, machine-learning algorithm in clinical decision making.

Study Design and Setting: Using a database containing comprehensive health assessment information (the interRAI-HC) on home care clients ($N = 24{,}724$) from eight community-care regions in Ontario, Canada, we compare the performance of the *K*-nearest neighbor (KNN) algorithm and a Clinical Assessment Protocol (the "ADLCAP") currently used to predict rehabilitation potential. For our purposes, we define a patient as having rehabilitation potential if the patient had functional improvement or remained at home over a follow-up period of approximately 1 year.

Results: The KNN algorithm has a lower false positive rate in all but one of the eight regions in the sample, and lower false negative rates in all regions. Compared using likelihood ratio statistics, KNN is uniformly more informative than the ADLCAP.

Conclusion: This article illustrates the potential for a machine-learning algorithm to enhance clinical decision making.

Keywords: Bayes' theorem; Clinical decision making; Diagnostic likelihood ratio; InterRAI; Machine learning; Rehabilitation

1. Introduction

There is increasing interest in applying sophisticated computer-modeling and statistical analysis techniques in health care [1,2]. With the expanded use of standardized information systems in many parts of the health system, there is great potential for enhanced use of these data to inform clinical decision making and health system planning.

A major research priority has been identified related to the need for improved methods of selecting older patients who are most likely to benefit from rehabilitation [3]. More accurate targeting of rehabilitation services would result in more efficient use of health care resources, reduction in health care costs resulting from the avoidance of functional decline, and improved quality of life for many older persons.

This issue is particularly important in that many older persons who would benefit from rehabilitation do not receive it [4]. Inadequate provision of rehabilitation is in part a reflection of resource constraints, but also reflects shortcomings in the way the health system manages patient information. For example, the lack of comprehensive standardized assessments often leads to under-detection of health care needs. Also, data are usually not used to their fullest potential because they tend to be considered (1) one item at a time or (2) with simple algorithms that do not allow for the detection of more complex, interactive effects. In many care settings, this information gap could be addressed through better use of existing health information systems.

1.1. The Resident Assessment Instrument−Home Care

The Resident Assessment Instrument−Home Care (RAI-HC) is a comprehensive assessment and problem

* Corresponding author. Tel.: 519-888-4567 ext. 37559; fax: 519-725-0784.

E-mail address: m3zhu@uwaterloo.ca (M. Zhu) or wchen@uwaterloo.ca (W. Chen) or hirdes@uwaterloo.ca (J.P. Hirdes) or stolee@uwaterloo.ca (P. Stolee).

0895-4356/07/$ − see front matter © 2007 Elsevier Inc. All rights reserved.
doi: 10.1016/j.jclinepi.2007.06.001

identification system developed by an international consortium of researchers (interRAI; http://www.interrai.org/). It is a part of a growing family of assessment tools developed for use in many health care settings for care and service planning, resource allocation, outcome measurement, and quality improvement [5].

The RAI-HC is mandated for use in all of Ontario's Community Care Access Centres (CCACs) for longer-term (more than 60 days) clients. CCACs coordinate access to home care services and long-term care placement in Ontario. This instrument is now also used or being implemented in numerous other jurisdictions in North America, Europe, and the Pacific Rim.

Assessment items include personal items, referral information, cognition, communication and hearing, vision, mood and behavior, informal support services, physical functioning, continence, disease diagnoses, preventive health measures, nutrition/hydration status, oral health, skin condition, environmental assessment, and service utilization. A number of standard interRAI scales have been derived from the assessment items, measuring such important health domains as cognition, depression, and ability to perform activities of daily living. Clinical Assessment Protocols (CAPs) are triggered when specified combinations of assessment items suggest that specific problems or risks are present and warrant further investigation [6−8].

1.2. Objectives

A number of recent works [9,10] have applied machine-learning techniques to predicting various rehabilitation outcomes. Although some of these results have been promising [9], others have been equivocal [10]. In this study, we investigate whether an automatic, data-driven, machine-learning algorithm is capable of accurately assessing a patient's rehabilitation potential. Most clinical practitioners are still largely skeptical that such assessment can be carried out by a "blind" algorithm at all. As such, we have chosen to start by focusing on one of the simplest machine-learning algorithms available, the K-nearest neighbor (KNN) algorithm [11], because it is analogous to clinical reasoning (see Section 2.2) and may therefore be more readily accepted by clinicians. The most relevant CAP for predicting rehabilitation potential is called the ADLCAP, where "ADL" stands for "activities of daily living." In evaluating the usefulness of KNN, we will therefore compare the predictions made by KNN against those made by the existing ADLCAP.

2. The KNN algorithm

2.1. Description

Suppose we have a database consisting of a total of n observations, (x_i, y_i), for $i = 1, 2, \ldots, n$. For our study, x_i is a vector of covariates and y_i is a binary outcome. In our case, $y_i = 1$ means patient i has rehabilitation potential. The database is called the *training set* for the KNN algorithm. Given any two observations, x_i and x_j, let $s(x_i, x_j)$ be a measure of their similarity based on the covariates. To predict the response for a new observation x_0 with the KNN algorithm, we first identify K observations in the training set that are most similar to x_0; they form the set of KNNs of x_0, denoted by $N(x_0, K)$. We then estimate the probability that $y_0 = 1$ by the average responses of these KNNs and predict the response to be one if the estimated probability exceeds a certain threshold c.

Hence, to implement the KNN algorithm, three ingredients must be specified a priori: the similarity measure, $s(x_i, x_j)$; the number of neighbors, K; and the decision threshold, c. The exact specifications of these ingredients for our study are given and justified below in Section 4. A more precise and mathematical description of the KNN algorithm is given in Appendix A.

2.2. KNN as an artificial "super expert"

To a certain degree, it can be argued that physicians also rely on an implicit KNN algorithm to make clinical decisions. A physician's clinical decision is undoubtedly influenced by his or her past clinical experiences. For example, a physician will likely recommend a particular treatment program to a new patient if the new patient's clinical profile matches those patients who have been successfully treated by the physician in the past with the same program. Hence, a physician's past patients can be regarded as the training set. Matching the clinical profile of a new patient to those of his or her past patients is similar to finding a number of nearest neighbors from the training set. In this sense, we can think of the KNN algorithm as an artificial "super expert" who has had the "experience" of "treating" virtually every patient recorded in the database and can, therefore, use this extensive "clinical experience" to make informed and intelligent decisions.

2.3. Software

The software we use for the KNN algorithm is a function called knn in the statistical package, R [12].

3. Data

Generally speaking, whether a patient has true rehabilitation potential or not is unknown, which is precisely why we are interested in different ways of predicting it. To apply machine-learning techniques and evaluate the accuracy of different prediction methods, however, we must rely on patients whose true rehabilitation potential can be reliably assessed by some other means.

For this investigation, we use RAI-HC data on 24,724 patients from eight Ontario CCACs. For these patients, their RAI-HC data have been linked to health-service utilization data, including long-term care admissions and mortality, which we can use to assess the patients' true rehabilitation potential. In this study, we define a patient as having true rehabilitation potential ($y = 1$) if (1) there is an improvement in the patient's ADL functioning (measured using the interRAI ADL long form [13]) over a follow-up period of approximately 1 year; or if (2) the patient remains at home at the end of the treatment program. In our data set, 6,567 patients are so defined as having true rehabilitation potential. Other disposition outcomes include discharge to a nursing home, or death, which could be considered indications of rehabilitation failure.

Therefore, in our study, we are asking the following question: Which method can better predict the two positive outcomes, (1) and (2), defined above, KNN or the existing ADLCAP?

4. Method

4.1. Covariates x

To make a fair and objective comparison between KNN and ADLCAP, we use exactly the same covariates in KNN as the ones used by ADLCAP. These are covariates related to various aspects of physical functioning, comprehension, health status indicators, and functional potential; see Table 1.

The ADLCAP is derived using a number of nested if-then statements that use different combinations of these variables as conditions [14]. In descriptive terms, the ADLCAP is triggered if the patient is unable to perform two or more of the ''activities of daily living'' items (h2a to h2j in Table 1); if the patient is able to understand others (c2); *and* if any one of the conditions described by the other covariates is present (p6 to h7c in Table 1). Although there is no explicit weighting of the items, the protocol implies a particular importance for the ability to understand others (c3)—considered necessary for the success of a rehabilitative program—as this is the one single item that must always be present for the ADLCAP to be triggered.

Table 1
Items used by ADLCAP as covariates

Item	Brief description
h2a	Mobility in bed
h2b	Transfer
h2c	Locomotion in home
h2d	Locomotion outside of home
h2e	Dressing upper body
h2f	Dressing lower body
h2g	Eating
h2h	Toilet use
h2i	Personal hygiene
h2j	Bathing
c3	Ability to understand others
p6	Overall change in care needs
h3	ADL decline
k8b	Condition unstable
k8c	Flare-up of chronic problem
k8d	Treatments changed in last 30 days
h7a	Client believes he/she is capable of increased functional independence
h7b	Caregiver believes client is capable of increased functional independence
h7c	Good prospects of recovery from current disease

4.2. Training set for KNN

We use KNN to make predictions on the eight regional CCAC data sets one by one. When predicting a particular region, we take a random sample of 2,500 clients from the other seven data sets and use it as the training set. This strategy automatically allows KNN to avoid using one's own data to predict itself (and thereby creating a bias toward better prediction).

4.3. Similarity measure $s(x_i, x_j)$

Again, to make our comparison fair and objective, we define and use a similarity measure $s(x_i, x_j)$ so that KNN and ADLCAP not only use exactly the same covariates, but also interpret these covariates in exactly the same manner.

Suppose there are a total of p covariates. We define the similarity $s(x_i, x_j)$ to be the total number of covariates that ADLCAP interprets as identical for x_i and x_j, that is,

$$s(x_i, x_j) = \#\{x_{id} \text{ regarded as identical to } x_{jd} \text{ by the ADLCAP;} \quad d = 1, 2, \ldots, p\}.$$

What does it mean to say a covariate is ''regarded as identical'' for two patients by ADLCAP? This is best explained with a prototypical example. The ADLCAP consists of a number of ''if-then'' statements. For instance, the variable ''h2a'' (mobility in bed) is treated in the following way [14]:

```
if (h2a = 2, 3, 4, 5, 6 or 8)
then do A;
else do B.
```

Thus, if $h2a_i = 2$ and $h2a_j = 6$, then patients i and j are regarded to have identical values for the covariate ''h2a,'' whereas if $h2a_i = 2$ and $h2a_j = 0$, patients i and j are regarded to have different values for the covariate ''h2a.''

4.4. Number of neighbors K

The most important parameter in the KNN algorithm is K, the number of neighbors. The choice of K involves an important trade-off and it must be selected with care. If K is too large, then some of the neighbors used to make predictions will no longer be similar to the one being predicted; this will bias the prediction. On the other hand, if K is too small, then not enough information is used to make the prediction; this will cause the prediction to be unstable. The optimal K is the one that best balances this trade-off [15]. Typically, the optimal K is selected empirically by using a procedure called cross-validation on the training set. Cross-validation is a standard procedure in machine learning, for example, Hastie et al. [15]; its details are not directly relevant to our study here and hence omitted. The software we use to perform cross-validation for KNN is the function knn.cv from the statistical package, R [12].

Figure 1 shows the cross-validated estimates of KNN's overall error rate on each of the eight training sets as K varies. For example, training set 1 is used by KNN to make predictions for region 1, so it consists of data from all other regions except region 1. Generally speaking, the overall error rate drops as K is increased and levels off at around $K = 20$. In few cases, we can see that the overall error rate starts to increase again as K is increased further, a clear indication of the trade-off discussed in the previous paragraph. Based on these results, we choose K to be 20.

4.5. Decision threshold c

We choose the decision threshold c to be the overall percentage of patients in our data set with *true* rehabilitation potential. As stated in Section 3, it is possible to reliably assess the true rehabilitation potential of patients in our data set by means other than the ADLCAP or the KNN algorithm. In particular, patients who showed functional improvement or remained at home over a 12-month follow-up period are considered to have true rehabilitation potential. Across the eight CCAC data sets, this percentage is 26.56% (6,567/24,724 ≈ 26.56%; see Section 3).

Put more explicitly, our decision threshold is that 26.56% or more of a patient's 20 nearest neighbors must have had positive rehabilitation outcomes (as defined in Section 3) for the patient to be assigned positive rehabilitation potential by the KNN algorithm. Determined by the actual proportion of patients who had positive rehabilitation outcomes in our data set, the threshold of 26.56% is seen to more closely emulate the actual clinical profile of these patients than any other (necessarily more arbitrary) threshold choices.

The intuition here is as follows: Suppose Dr. Kevin N. Newman (initial K.N.N.) is our super expert. Dr. Newman knows from his extensive past experience (all patients in our data set) that only 26.56% of his patients could really benefit from rehabilitation. He is examining a new patient right now and the new patient looks a lot like these 20 other patients he once saw. Dr. Newman also knows that, among these 20 patients from the past, more than 26.56% of them benefited from rehabilitation. Therefore, he concludes this new patient should have a high chance of being able to benefit also.

4.6. Treatment of ties

It is possible that a number of observations from the training set are tied for the Kth position in terms of their similarities to x_0, the new observation being evaluated/assessed/predicted. Under such circumstances, there are two standard options: (1) randomly pick one of them so that the set $N(x_0, K)$ contains exactly K items or (2) use all of them, allowing the set $N(x_0, K)$ to contain more than K items (even though this would seem to contradict the name K.N.N.). The knn function in R allows the user to choose between these two options. We have chosen option (2), but this choice is arbitrary and does not make a tangible difference in the final results we present below.

The main reason why we have chosen option (2) is because, with option (1), a small number of predictions could

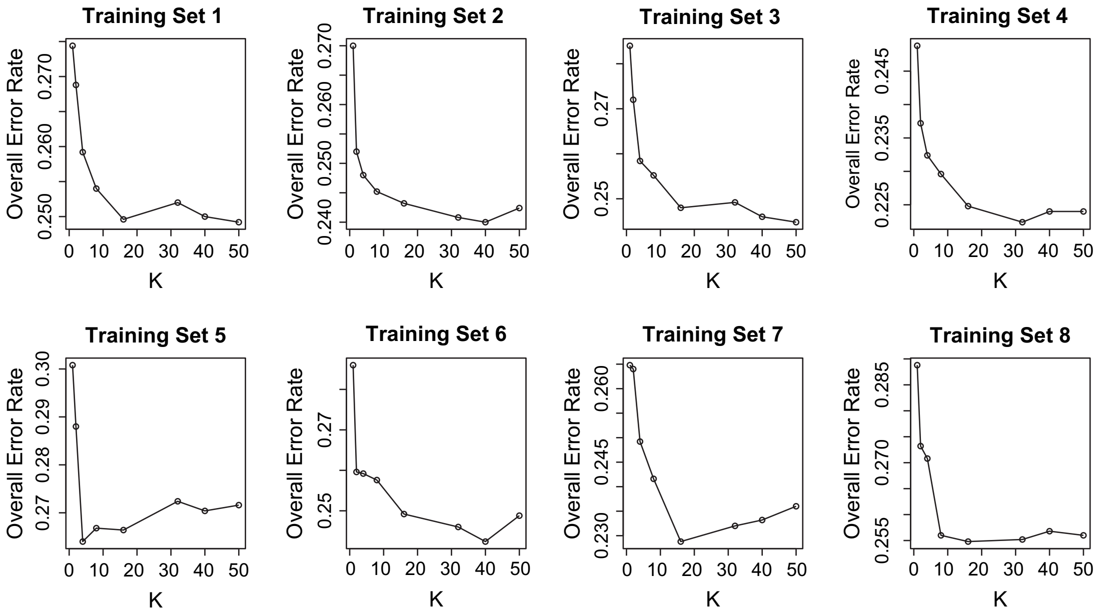

Fig. 1. Cross-validated overall error rate vs. K.

change (e.g., from a zero to a one) when we repeat the analysis on the same data due to the random selection step. This causes a bit of practical inconvenience because we would not be able to reproduce our results (e.g., Tables 3 and 4) exactly even if we use exactly the same training sets. We must add, however, that the resulting variation is minor and inconsequential; only a few numbers would change in the fourth decimal place.

4.7. Evaluation criteria

We are now faced with two competing methods to evaluate/assess the rehabilitation potential of a patient, ADLCAP and KNN. Both give a binary prediction $\hat{y}$. We evaluate the accuracy of these two methods using four different criteria: the false positive (FP) rate, the false negative (FN) rate, the positive diagnostic likelihood ratio (DLR+), and the negative diagnostic likelihood ratio (DLR−). Their respective definitions are as follows:

$$\begin{aligned}\mathrm{FP} &= P(\hat{y}=1|y=0),\\ \mathrm{FN} &= P(\hat{y}=0|y=1),\\ \mathrm{DLR}+ &= \frac{P(\hat{y}=1|y=1)}{P(\hat{y}=1|y=0)},\\ \mathrm{DLR}- &= \frac{P(\hat{y}=0|y=1)}{P(\hat{y}=0|y=0)},\end{aligned}$$

where y is the true status and $\hat{y}$ is the predicted status. These are standard criteria [16] and widely used in clinical epidemiology.

The FP and FN criteria are more intuitive, corresponding, respectively, to the two types of errors one can make in a binary prediction, namely, calling a true zero a one (FP) and calling a true one a zero (FN).

The DLR+ and DLR− criteria are less intuitive but extremely useful, because they “quantify the change in the odds [of y] conferred by knowledge of [$\hat{y}$]” or “the increase in knowledge about [the true status y] gained through [the prediction $\hat{y}$]” [16]. To see this, start with the prior odds and the posterior odds, defined as

$$\text{prior odds} = \frac{P(y=1)}{P(y=0)} \quad \text{and}$$

$$\text{posterior odds} = \frac{P(y=1|\hat{y})}{P(y=0|\hat{y})}.$$

By a simple application of Bayes’ theorem [17], it can be shown (see Appendix B) that

$$\text{posterior odds}(\hat{y}=1) = (\mathrm{DLR}+) \times \text{prior odds}, \qquad (1)$$

$$\text{posterior odds}(\hat{y}=0) = (\mathrm{DLR}-) \times \text{prior odds}. \qquad (2)$$

Table 2
A confusion matrix

	$y=1$	$y=0$
$\hat{y}$=1	a	b
$\hat{y}$=0	c	d

The entries a, b, c, and d are counts of the number of observations falling into each of the four cells.

Therefore, DLR+ can be interpreted as the factor by which a prediction of $\hat{y}$=1 can increase the prior odds and DLR−, the factor by which a prediction of $\hat{y}$=0 can decrease the prior odds. Clearly, we would expect an informative prediction method to have DLR+ > 1 and DLR− < 1. In fact, given two prediction methods, A and B, A can be said to be more informative than B if DLR+ (A) > DLR+ (B) and DLR− (A) < DLR− (B).

Partly due to these rather sophisticated interpretations associated with them, DLR+ and DLR− have been gaining popularity in the last two decades [16,18,19]. In terms of the 2 × 2 confusion matrix (Table 2), these criteria can be calculated as

$$\begin{aligned}\mathrm{FP} &= b/(b+d),\\ \mathrm{FN} &= c/(a+c),\\ \mathrm{DLR}+ &= \frac{a/(a+c)}{b/(b+d)},\\ \mathrm{DLR}- &= \frac{c/(a+c)}{d/(b+d)}.\end{aligned}$$

The c statistic is another commonly used evaluation criterion [20]. To predict a binary outcome y, one often estimates the probability $P(y=1)$ first, and then threshold the probability at some level (e.g., 50%). As the threshold changes, the final binary prediction and the resulting FP and FP rates will also change. The c statistic depends on the entire profile of FP and FP rates as we vary the threshold. Because the ADLCAP produces just the final binary prediction, it is impossible to calculate a c statistic for it. Therefore, this criterion is not used in our analysis.

4.8. Ethics approval

This study has received ethics clearance from the Office of Research Ethics at the University of Waterloo.

5. Results

Tables 3 and 4 contain the results comparing KNN against ADLCAP using the four different criteria outlined in Section

Table 3
Comparative results: FP rate (False+) and FN rate (False−)

Region	False+		False−	
ID	CAP	KNN	CAP	KNN
1	0.2957	0.3385	0.6498	0.3628
2	0.3085	0.3067	0.6162	0.3838
3	0.3211	0.2733	0.6330	0.4967
4	0.3569	0.3011	0.6451	0.3523
5	0.2657	0.1843	0.6684	0.5263
6	0.3754	0.2438	0.6222	0.4137
7	0.4310	0.2763	0.5896	0.3676
8	0.3730	0.2783	0.6154	0.4218

Abbreviation: “CAP” refers to “ADLCAP”.

Table 4
Comparative results: DLRs

Region	DLR+		DLR−	
ID	CAP	KNN	CAP	KNN
1	1.1841	1.8826	0.9227	0.5484
2	1.2442	2.0088	0.8911	0.5537
3	1.1431	1.8415	0.9323	0.6835
4	0.9944	2.1511	1.0031	0.5040
5	1.2479	2.5704	0.9103	0.6452
6	1.0062	2.4049	0.9963	0.5470
7	0.9521	2.2882	1.0363	0.5080
8	1.0311	2.0775	0.9815	0.5844

Abbreviations: DLR+, positive diagnostic likelihood ratio; DLR−, negative diagnostic likelihood ratio; "CAP" refers to "ADLCAP."

4.7. Other than in region 1 where it has a slightly higher FP rate, KNN makes better predictions than ADLCAP in all other regions, having both lower FP and lower FN (Table 3).

More significantly, on the DLR+ and DLR− scales (Table 4), KNN emerges as a uniformly more informative method than ADLCAP for assessing rehabilitation potential in all regions. As explained in Section 4.7, all informative prediction methods are expected to have DLR+ $>$ 1 and DLR− $<$ 1. Here, we notice that, in regions 4 and 7, the ADLCAP actually has a DLR+ slightly less than 1 and a DLR− slightly bigger than 1.

6. Discussion

This is a comparative study; our goal was to investigate whether KNN could be a more effective algorithm than the CAP currently used for assessing and predicting rehabilitation potential (ADLCAP). In order for the comparison to be objective, we have only used covariates that are also used in the ADLCAP. We have also defined a highly specialized similarity measure so that the covariates are interpreted in the same way as in the ADLCAP. In other words, we have tried not to give KNN any extra advantage. Even within these constraints, we have shown that a data-driven algorithm such as KNN can be more informative than the ADLCAP.

In reality, of course, we are by no means required to restrict ourselves to this particular subset of covariates or this particular type of similarity measure. A second advantage we saw in applying these constraints, however, was that the KNN algorithm would only be making use of variables that had been identified clinically as relevant to rehabilitation potential. We recognize that machine-learning algorithms can be seen as a "black box" by clinicians. Restricting our analyses to the same variables as are used in accepted clinical protocols could enhance the acceptability of this approach. As discussed earlier in the article, we also saw the KNN algorithm as being analogous to clinical reasoning and therefore relatively easy to justify to clinicians and other decision makers.

A limitation of this study is that, while we were assessing rehabilitation potential using an accepted set of relevant covariates, we were obviously not in a position to determine whether the positive (or negative) outcomes of the home care clients in this sample were attributable to the provision (or the absence) of any rehabilitative therapy. But the results suggest that both approaches, and particularly the ADLCAP, are conservative—many more patients had positive outcomes than were predicted. The interRAI consortium is currently in the process of refining the CAPs for its entire suite of instruments; our results could provide some guidance for that exercise. For example, results from KNN provide a performance benchmark against which clinical protocols could be measured—clinical judgment could be used to identify where the CAP rules could be changed to achieve better prediction results approaching those from the KNN.

We are continuing to explore these issues, as well as working to identify better approaches to applying the KNN algorithm and other machine-learning techniques in clinical decision making. For example, a limitation of the KNN algorithm is that, to predict new cases, the entire data set has to be stored in memory and accessible to the person trying to make the prediction. To this end, a possible strategy is to find a small number of prototype cases from the data set and apply nearest-neighbor-type algorithms on the prototypes alone. The resulting algorithm will also become more interpretable (and hence less of a "black box") if the number of prototypes is relatively small.

This study illustrates the potential for a machine-learning algorithm to enhance clinical decision making. As use of computerized health information systems, such as those based on the interRAI instruments, becomes more widespread, there is great potential to use these algorithms to direct therapy and services at those patients most likely to benefit. We recommend greater exploration by the interRAI consortium in regard to applications of these techniques in clinical assessment and care planning based on interRAI data. More generally, our work is relevant to all those working to make better use of standardized heath information in clinical decision making and service planning.

Acknowledgments

This research is partially supported by funding from the Canadian Institutes of Health Research Institute of Musculoskeletal Health and Arthritis, the Ontario Rehabilitation Research Network, the Natural Science and Engineering Research Council of Canada, the Mathematics of Information Technology And Complex Systems network, and the Canadian National Program on Complex Data Structures.

Appendix A

Detailed description of KNN

In this appendix, we give a more precise and mathematical description of the KNN algorithm. Given the covariate vector of a new observation, x_0, the goal is to predict its response, y_0. For every observation x_i in the training set, let $s_i = s(x_0, x_i)$ be its similarity to x_0. These similarities can be ordered. Denote the ordered similarities with $s_{(i)}$, that is, $s_{(1)} \geqslant s_{(2)} \geqslant \ldots \geqslant s_{(n)}$. In other words, if $s_j = s_{(k)}$, it means x_j is the kth most similar observation in the training set to x_0. The set of the KNNs of x_0, $N(x_0, K)$, can then be defined as all observations whose similarities to x_0 are at least $s_{(K)}$, that is, $N(x_0, K) = \{x_i : s_i \geqslant s_{(K)}\}$. The KNN algorithm then estimates the probability that $y_0 = 1$ with

$$\hat{p} = \frac{\sum_{x_i \in N(x_0,K)} y_i}{|N(x_0,K)|}$$

where $|N(x_0, K)|$ is the size of (or number of items contained in) the set $N(x_0, K)$. This is usually equal to K exactly, but may exceed K depending on how ties are treated (see Section 4). The response is then predicted to be one if $\hat{p} \geqslant c$, where c is a prespecified threshold parameter.

Appendix B

Derivation of equation (1)

In this appendix, we derive equation (1), given in Section 4.7; equation (2) can be derived in exactly the same manner. These derivations are standard [16], but we include them here for convenience. By definition,

$$\text{posterior odds}(\hat{y} = 1) = \frac{P(y = 1|\hat{y} = 1)}{P(y = 0|\hat{y} = 1)}.$$

Apply Bayes' theorem [17] to both the denominator and the numerator, and we get

$$\begin{aligned}\text{posterior odds}(\hat{y} = 1) &= \frac{P(y = 1|\hat{y} = 1)}{P(y = 0|\hat{y} = 1)} \\ &= \frac{\dfrac{P(\hat{y} = 1|y = 1)P(y = 1)}{P(\hat{y} = 1)}}{\dfrac{P(\hat{y} = 1|y = 0)P(y = 0)}{P(\hat{y} = 1)}} \\ &= \frac{P(\hat{y} = 1|y = 1)}{P(\hat{y} = 1|y = 0)} \times \frac{P(y = 1)}{P(y = 0)} \\ &= (\text{DLR+}) \times \text{prior odds}.\end{aligned}$$

References

[1] Tremblay M. Predictive health: policy for predictive modelling and long-term health conditions. Report prepared for Department of Health, England. London: Tremblay Consulting; 2005.

[2] Mitnitski AB, Mogilner AJ, Graham JE, Rockwood K. Techniques for knowledge discovery in existing biomedical databases: estimation of individual aging effects in cognition in relation to dementia. J Clin Epidemiol 2003;56:116–23.

[3] Stolee P, Borrie MJ, Cook S, Hollomby Jthe participants of the Canadian Consensus Workshop on Geriatric Rehabilitation. A research agenda for geriatric rehabilitation: the Canadian consensus. Geriatr Today J Can Geriatr Soc 2004;7:38–42.

[4] Hirdes JP, Fires BE, Morris JN, Ikegami N, Zimmerman D, Dalby DM, et al. Home Care Quality Indicators (HCQIs) based on the MDS-HC. Gerontologist 2004;44:665–79.

[5] Hirdes JP, Fries BE, Morris J, Steel K, Mor V, Frijters DH, et al. Integrated health information systems based on the RAI/MDS series of instruments. Health Manage Forum 1999;12:30–40.

[6] Morris JN, Fries BE, Steel K, Ikegami N, Bernabei R, Carpenter GI, et al. Comprehensive clinical assessment in community setting: applicability of the MDS-HC. J Am Geriatr Soc 1997;45:1017–24.

[7] Landi F, Tua E, Onder G, Carrara B, Sgadari A, Rinaldi C, et al. Minimum data set for home care: a valid instrument to assess frail older people living in the community. Med Care 2000;38:1184–90.

[8] Diwan S, Shugarman LR, Fries BE. Problem identification and care plan responses in a home and community-based services program. Med Care 2004;23:193–211.

[9] Tam SF, Cheing GLY, Hui-Chan SWY. Predicting osteoarthritic knee rehabilitation outcome by using a prediction model developed by data mining techniques. Int J Rehabil Res 2004;27:65–9.

[10] Ottenbacher KJ, Linn RT, Smith PM, Illig SB, Mancuso M, Granger CV. Comparison of logistic regression and neural network analysis applied to predicting living setting after hip fracture. Ann Epidemiol 2004;14:551–9.

[11] Cover TM, Hart PE. Nearest neighbor pattern classification. IEEE T Inform Theory 1967;13(1):21–7.

[12] R Development Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, ISBN 3-900051-07-0; 2006. Available at http://www.R-project.org.

[13] Morris JN, Fries BE, Morris SA. Scaling ADLs within the MDS. J Gerontol Med Sci 1999;54:M546–53.

[14] Morris JN, Fries BE, Steel K, Ikegami N, Bernabei R. Primer on Use of the Minimum Data Set-Home Care (MDS-HC) Version 2.0© and the Client Assessment Protocols (CAPs). Boston: Hebrew Rehabilitation Center for Aged; 1999.

[15] Hastie TJ, Tibshirani RJ, Friedman JH. The elements of statistical learning: data-mining, inference and prediction. New York: Springer Verlag; 2001.

[16] Pepe MS. The statistical evaluation of medical tests for classification and prediction. New York: Oxford University Press; 2003.

[17] Bayes T. An essay towards solving a problem in the doctrine of chances. Phil Trans Roy Soc 1763;53:370–418.

[18] Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical epidemiology: a basic science for clinical medicine. 2nd edition. Boston: Little, Brown and Company; 1991.

[19] Greenhalgh T. How to read a paper: papers that report diagnostic or screening tests. Br Med J 1997;315:540–3.

[20] Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1998;4: 361–87.