A Simple Method for Heuristic Modeling of Expert Knowledge in Chronic Disease: Identification of Prognostic Subgroups in Rheumatology

Örjan DAHLSTRÖM a, b, 1, Toomas TIMPKA c, d, Ursula HASS c, Thomas SKOGH e and Ingrid THYBERG e

a The Swedish Institute for Disability Research, Linköping and Örebro University, Sweden
b Department of Behavioural Sciences and Learning, Linköping University, Sweden
c Department of Health and Society, Linköping University, Sweden
d Department of Computer and Information Science, Linköping University, Sweden
e Department of Molecular and Clinical Medicine, Linköping University, Sweden

Abstract. Identification of prognostic subgroups is of key clinical interest at the early stages of chronic disease. The aim of this study is to examine whether representation of physicians’ expert knowledge in a simple heuristic model can improve data mining methods in prognostic assessments of patients with rheumatoid arthritis (RA). Five rheumatology consultants’ experiences of clinical data patterns among RA patients, as distinguished from healthy reference populations, were formally represented in a simple heuristic model. The model was used in K-mean-clustering to determine prognostic subgroups. Cross-sectional validation using physician’s global assessment scores indicated that the simple heuristic model performed better than crude data made in identification of prognostic subgroups of RA patients. A simple heuristic model of experts’ knowledge was found useful for semi-automatic data mining in the chronic disease setting. Further studies using categorical baseline data and prospective outcome variables are warranted and will be examined in the Swedish TIRA-program.

Keywords. Knowledge engineering, Clinical Decision Support Systems, Semi-automated Data Mining, Rheumatoid Arthritis, Mathematical models in medicine.

Introduction

Chronic diseases, such as cardiovascular disease, diabetes, and rheumatic diseases constitute a large part of the disease burden in western countries [1]. Early identification of patients with chronic disease at risk of progression is essential, especially when early intervention can induce remission or retard the disease process.

1 Corresponding Author: Örjan Dahlström, Department of Behavioural Sciences and Learning, SE-581 83 Linköping, Sweden. E-mail: orjan.dahlstrom@liu.se.
The prevalence of rheumatoid arthritis (RA) in an adult Swedish population is about 5-7 cases per 1000 individuals [2] and the yearly incidence is 24/100,000 [3]. Identifying subgroups of patients with a less favorable prognosis at an early stage is therefore both a key clinical and public health priority.

Much of clinical decision support research has addressed the establishment of a clinical diagnosis during a short-term care episode. This approach has often been developed using comparisons between physicians’ performances as a baseline, implying a replacement of the decision-maker rather than integration of the decision support system as a component in the actual clinical care environment. Such an approach has less relevance in the chronic disease context, where the diagnosis is already known. A more appropriate strategy here is that the decision support supplies the decision-maker with contextual information at different stages in the clinical process. This strategy is compatible with theories of distributed cognition [4], in which both human and machine agents are considered as being integral parts of a functional cognitive system, that recently have been applied in clinical environments [5].

A specific type of decision support is assistance with knowledge discovery in databases (KDD) [6]. Integration of background knowledge into discovery methods can improve the quality of these processes. This tactic has been used for, e.g. identification of new clinically interesting subgroups [7]. However, putting the medical expert in the center of the KDD process, described as active mining, can be time-consuming and also expensive. Therefore, semi-automated active mining has been introduced. This method has shown to be promising with regard to automated discovery from an incomplete data set in early detection of patient groups at risk for coronary heart disease [8].

The aim of this study is to examine whether representation of physicians’ expert knowledge in a simple heuristic model can be used to improve data mining methods in prognostic assessments of RA patients. The starting point for the heuristic modeling is that the clinical test values for patients with a specific chronic disease differ from those of healthy reference populations.

1. Materials and Methods

The research was performed in three steps. In a knowledge engineering step, physicians’ experiences from the clinical use of four variables included in the evaluation of RA patients were elicited. In the second step, these data were analyzed and represented in a model that was implemented in an algorithm for the determination of prognostic groups. For comparison, the same algorithm was also used to develop groups based on the crude data set. In the final step, prognostic group validation, the model was validated by comparing physicians’ global assessment of disease activity scores (PGA (scores 0–4, where 0 corresponds to no activity and 4 represents high activity)) for the patient groups identified with and without the model.

Knowledge Engineering

The knowledge engineering was based on data from a Swedish ‘early arthritis’ database [9]. This resource has been built in a prospective multi-center study (TIRA) where 320 patients diagnosed with RA were diagnosed early and included during 1996 – 1998. The patients fulfilled ≥4/7 RA classification criteria (95%), as defined by the
American College of Rheumatology 1987, or exhibited morning stiffness ≥60 minutes, symmetrical arthritis, and arthritis of small joints (5%). Although several prognostic indicators can be suggested, the interval-/ratio variables in the database were especially interesting for this study, because they could be used in the standard method of K-mean clustering. The set of variables used here included the erythrocyte sedimentation rate (ESR), C-reactive protein (CRP), the ‘number of swollen joints’ and the ‘number of tender joints’.

Five consultants in rheumatology were interviewed. The interviews were held in a semi-structured manner and aimed at investigating the physicians’ perspectives on important variables when making prognostic judgments about patients with RA. The consultants had the opinion that the variables ‘number of swollen joints’ and ‘number of tender joints’ were not sufficient stand-alone indicators of the patient’s status. These were therefore omitted from further analysis. ESR and CRP, however, were both considered to be important and relevant indicators of patients’ status. We wanted the modeling of their expert knowledge to be as simple as possible to interpret and therefore used the notions values. We therefore modeled their views on ESR and CRP data by using the values they defined as low, elevated, and definitely high as limits for categorizations of ESR and CRP. The consultants’ perspectives on what are considered as low values were almost identical (ESR ≤ 12 mm for men and ESR ≤ 20 mm for women, CRP ≤ 10 mg/L for both men and women), with the exception of some minor differences regarding low values for ESR. Opinions of what are to be considered as elevated or definitely high values differed only slightly (an elevated value is roughly ESR = 60 mm and CRP = 70 mg/L, while ESR = 100 mm and CRP = 100 mg/L are definitely high values).

To give ESR and CRP equal influence we transformed the ESR and CRP values to the interval [0, 1]. We estimated 90% of the values to be represented by values ranging from low to elevated (0, .9) by a linear transformation. Values ranging from elevated to definitely high were transformed into the interval [.9, 1) in a decreasing manner using Eq. (1). Values considered as low were transformed to 0 and values considered as definitely high were transformed to 1.

\[
f(x) = p + (1 - p) \cdot \left[ 1 - \left( 1 - \frac{x - x_0}{x_2 - x_1} \right)^{\frac{p}{1 - p}} \right]
\]

In Eq. (1), x is the variable for which the consultants’ experience is being modeled (ESR or CRP), f is the modeled value, x0 is the ‘lower limit value’ for which all lower values are considered as low, x1 is the ‘first upper limit value’, the point at which higher values are considered as elevated, x2 is the ‘final upper limit value’ above which all higher values are considered definitely high, p is the width of the interval which (x0, x1) is transformed to.

The ‘limit values’ used in Eq. (1) for ESR and CRP in the knowledge engineering step are shown in Table 1.
Table 1. Values for model limits $x_0$, $x_1$, $x_2$, and $p$ determined in the knowledge engineering step when modeling consultants' views of erythrocyte sedimentation rate (ESR) and C-reactive protein (CRP).

<table>
<thead>
<tr>
<th>Variable</th>
<th>$x_0$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESR (men)</td>
<td>12</td>
<td>60</td>
<td>100</td>
<td>.90</td>
</tr>
<tr>
<td>ESR (women)</td>
<td>20</td>
<td>60</td>
<td>100</td>
<td>.90</td>
</tr>
<tr>
<td>CRP (men and women)</td>
<td>10</td>
<td>70</td>
<td>100</td>
<td>.90</td>
</tr>
</tbody>
</table>

Determination of Prognostic Groups

To find different subgroups of patients, the method of K-mean clustering, was used on crude data, and then with two kinds of transformed data:

- Model A – Using the transformed data.
- Model B – Using the transformed data but with a predefined group of patients with values below the lower limit values ($x_0$) for both ESR and CRP.

To keep the results on a clinically comprehensive and not too complex level, the number of prognostic subgroups was limited to a maximum of four groups. The case with two subgroups was found trivial. Further cluster analyses were therefore restricted to three and four subgroups.

Prognostic Group Validation

In each set of the cluster analyses, the patients in the prognostic groups were compared with one another with respect to PGA. Subgroups, identified when the models were used, were compared with the results based on crude data to establish which method that best identified prognostic subgroups with respect to PGA using 95% confidence interval.

2. Results

Determination of Prognostic Groups

Running K-mean-clustering using three clusters resulted in a ‘low value group’ (low values for both ESR and CRP), a ‘medium value group’ (medium values for both ESR and CRP), and a ‘high value group’ (high values for both ESR and CRP). This is what could have been expected since there is a positive correlation (.66) between ESR and CRP. The same is true for the transformed values where the correlation between the modeled values for ESR and CRP are positive, varying between .66 and .68.

Running K-mean-clustering using four clusters made the situation more complex. As in the case of three clusters, a ‘low value group’ and a ‘high value group’ appeared. There were also two ‘medium value groups’. Despite minor differences in sizes and medium values for the clusters in the different model(s), the results were quite similar in this respect.
Table 2. Validation of derived models A and B against crude data using physicians’ global assessment scores (PGA) as proxy for patient outcome. Mean values (95% confidence interval) are displayed for each cluster (three and four clusters, respectively).

<table>
<thead>
<tr>
<th></th>
<th>Crude data</th>
<th>Model A</th>
<th>Model B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>PGA</td>
<td>N</td>
</tr>
<tr>
<td>3 Clusters</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 (low)</td>
<td>188</td>
<td>1.89</td>
<td>(1.78-2.01)</td>
</tr>
<tr>
<td>2 (medium)</td>
<td>70</td>
<td>2.10</td>
<td>(1.93-2.27)</td>
</tr>
<tr>
<td>3 (high)</td>
<td>21</td>
<td>2.33</td>
<td>(1.89-2.77)</td>
</tr>
<tr>
<td>4 Clusters</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 (low)</td>
<td>171</td>
<td>1.88</td>
<td>(1.76-2.00)</td>
</tr>
<tr>
<td>2 (medium)</td>
<td>36</td>
<td>2.19</td>
<td>(1.98-2.41)</td>
</tr>
<tr>
<td>3 (medium)</td>
<td>51</td>
<td>2.00</td>
<td>(1.77-2.23)</td>
</tr>
<tr>
<td>4 (high)</td>
<td>21</td>
<td>2.33</td>
<td>(1.89-2.77)</td>
</tr>
</tbody>
</table>

Prognostic Group Validation

In the case of three clusters, the three patient subgroups identified using model B on the basis of ESR and CRP differed significantly with regard to PGA scores (Table 2). Model A identified one significant difference. In comparison, the analyses based on the crude data identified no significant differences between any groups.

In the case of four clusters, no significant differences were found in the crude data model with regard to PGA. Model A and Model B, however, both allowed the identification of significantly different groups based on ESR and CRP (Model A: 1 significant difference, Model B: 3 significant differences).

3. Discussion

A knowledge engineering approach where heuristic expert knowledge was used to transform data by a simple procedure showed to be more efficient for identifying prognostic subgroups than K-mean clustering of crude data. The resulting subgroups passed the ‘objective’ gateway, which means that they included significant differences when using basic statistical criteria for cross-sectional evaluation, compared to the analysis of crude data [8]. The PGA score used for the evaluation is the variable in the TIRA database that was the closest proxy to long-term clinical outcomes. Passing the next, ‘subjective’, gateway will require that benefit is perceived when the procedure is used in a clinical action scenario.

Our approach to identifying prognostic subgroups differs from most other approaches in subgroup discovery [10]. These procedures mainly address the problem of large search spaces using constraint knowledge (restricting the search space), pattern knowledge (focusing the search) or ontological knowledge (applying weights to attributes) [7]. In other words, these methods are relevant mainly for larger datasets. Engineering knowledge in such settings aims usually to exclude uninteresting parts of the database. However, our main interest was not to mine data in large datasets, but to
examine heuristic modeling of expert knowledge in a form usable in standard data mining methods. We employed cluster analysis on a medium-sized dataset (n = 320), but our method can also be used in other methods and on larger datasets. For instance, our knowledge engineering modeling of expert knowledge is possible to integrate into other subgroup discovery approaches [7, 8] and into other general KDD approaches.

A main reason for using cluster analysis as the standard method for subgroup discovery is that understandability and interpretability of the generated models are of prime importance when viewing the knowledge discovery from the expert’s point of view. Results from cluster analysis are relatively simple to interpret and trace compared to the output from many other data mining approaches, e.g. support vector machines. The use of cluster analysis is an attempt to ‘open the black box’ of computations in data mining [11] for the clinician by choosing a method that is easy to understand. In the clinical context of chronic disease management, clinicians will be able to use our results in semi-automated data mining as part of the KDD processes. HEARTFAID is a comprehensive program that aims to improve the prognostic processes in the field of chronic cardiovascular disease [12]. The program has developed a similar architecture for an information infrastructure that is under development in the TIRA program in rheumatology. Our simple heuristic model fits into both of these contexts.

An important limitation of the model in its present form is that it is applicable only to variables on the interval-/ratio scale. The next step in the development of the model in the context of semi-automated data mining will be to include the examination of categorical data. Further research involving prospective outcome variables is warranted.

References