This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier’s archiving and manuscript policies are encouraged to visit:

http://www.elsevier.com/copyright
Case-based estimation of the risk of enterobiasis

Mare Remm\textsuperscript{a,b,*}, Kalle Remm\textsuperscript{b}

\textsuperscript{a} Tartu School of Health Care, Nooruse 9, 50411 Tartu, Estonia
\textsuperscript{b} Institute of Ecology and Earth Sciences, University of Tartu, Vanemuise 46, 51014 Tartu, Estonia

Received 20 May 2007; received in revised form 30 March 2008; accepted 6 April 2008

KEYWORDS
Case-based reasoning software; Risk estimation of enterobiasis; Nursery school children

Summary

Objective: To introduce an original case-based machine learning (ML) and prediction system \textit{Constud} and its application on tabular data for estimation of the risk of enterobiasis among nursery school children in Estonia.

Methods and materials: The system consists of a software application and a knowledge base of observation data, parameters, and results. The data were obtained from anal swabs for the diagnosis of enterobiasis, from questionnaires for children’s parents, observations in nursery schools and interviews with supervisors of the groups. The total number of studied children was 1905. Ten parallel ML processes were conducted to find the best set of weights for features and cases.

Results: The best goodness-of-fit according to the true skill statistic (TSS) was 0.381. Approximately equal fit can be reached using different sets of features. Cross-validation TSS of logit-regression and classification tree models was \(< 0.24. In addition to the higher prediction fit, \textit{Constud} is not sensitive to missing values of explanatory variables.

The overall prevalence of enterobiasis was 22.8%; the mean of risk estimations was 47.8%. The overestimation of the prevalence in risk calculations can be interpreted as an inefficacy of the single swab analysis, or may be due to the relative constancy of the risk compared to the lability of infection and the applied objective function.

Conclusions: In addition to the higher prediction fit, \textit{Constud} is not sensitive to missing values of explanatory variables. The main risk factors of enterobiasis among nursery school children were the child’s age, communication partners, habits, and cleanliness of rooms in the nursery school. Mixed age groups at nursery schools also enhance the risk.

\textcopyright{} 2008 Elsevier B.V. All rights reserved.

1. Introduction

Pinworm (\textit{Enterobius vermicularis}) is a contact helminth causing enterobiasis. Pinworm is the most common helminth in temperate climate countries.
The main risk factors are associated with indoor living conditions and close contact with other people, as pinworms spread mainly indoors directly from one human to another. The most commonly infected are children, particularly in groups. Although scarce and inconsistent statistical records indicate a decreasing trend of enterobiasis in Estonia, a special investigation found an infection level of 20% among nursery school children and a prevalence of up to 61% in nursery school groups [2]. A comparative investigation of risk factors and the relative role of the child’s age, personal habits and hygiene, children groups, family composition, and living conditions are needed in order to reduce the prevalence of enterobiasis.

Epidemiological risk factors constitute characteristics of persons and the environment that are related to a higher occurrence of a disease in a population. The epidemiological risk can be estimated as a risk associated with an individual or with a particular group of people, or alternatively, as describing certain characteristics of the environment or people.

The main method used to estimate the degree of influence of single risk factors is odds ratio (OR); the main methods to estimate the statistical significance of various risks analysed separately are χ²-test and Mann–Whitney U-test. OR is the ratio of exposure odds among cases to exposure odds among controls in a case–control study. The OR has been used in studies of enterobiasis, e.g., by John et al. [3] and Pezzani et al. [4]. The χ²-test is used to compare distributions of count data. It has been used in studies of the risk of enterobiasis [2,4,5]. The U-test is applicable to compare the expectations of numerical variables in two groups of unpaired observations. Such an epidemiological variable is, e.g., the level of IgG antibodies in blood [6].

Presumably, multivariate risk estimation should be more reliable than univariate estimations, which analyse every factor separately. The main branches of predictive computational multivariate methods are statistical modelling and artificial intelligence (AI) methods. Logistic regression is the most common method of statistical modelling used in estimations of risks, including epidemiological risks, because the interval of risk values is 0–1. Logistic regression has been used to estimate the risk of enterobiasis, e.g., by John et al. [3], Pezzani et al. [4] and Song et al. [5].

Problem-solving AI strategies can be either knowledge-intensive, also called rule-based reasoning, or search intensive, also called similarity-based reasoning or case-based reasoning (CBR) [7]. The rule-based technologies store their knowledge in rules or models; these methods include decision trees and artificial neural networks (ANNs), in principal also statistical modelling. Decision trees and ANN systems have been applied in medical diagnostics [8].

This paper presents a case study on the application of a case-based prediction system with which to estimate the level of risk and to compare the risk factors of enterobiasis among nursery school children in Estonia. A case-based estimation methodology is tried since it should be suitable for risk prediction according to its theoretical premises—no restrictions on the type and statistical distribution of variables, an abundance of observations exist, models and generalisations are not needed.

2. CBR and software Constud

2.1. CBR

CBR is defined as a multi-disciplinary research area and problem-solving paradigm that reuses previous experiences at a low level of generalisation to evaluate, interpret or solve a current new problem [7,9]. The knowledge of a CBR-system is stored in a case-base of raw observations; generalisations are represented as selected typical cases called exemplars. A classical case-based prediction system consists of exemplars, characteristics of exemplars—called features, rules for similarity estimation, feature weights, exemplar weights, retrieval indices and/or pre-classification, validation and adaption rules.

CBR has proved its strength in many weak-theory domains, in which a great number of single examples and case studies predominate over deduction, and in which large databases of previous cases exist, including medicine [10,11]. CBR is relatively easy to understand in terms of how the results are produced and which cases are used for them, and it can be effective even if the knowledge base is incomplete. A CBR-system is always ready to incorporate new observations and features and always ready to give hitherto the best predictions.

Although CBR in medicine is undergoing a revival [12], it has been applied relatively less frequently in medical research than other AI methods, according to Park et al. [8]. The approach should be promising, since clinical problem solving is traditionally highly case-specific and a lot of examples of its use exist. These include the estimation of life expectancy in a group of patients [13], clinical diagnoses and other medical decision-support systems [14–17], epidemic screening [18], predicting weekly incidences of influenza [19], and medical education [14,20]. In addition, computer vision and interpretation of medical images is methodologically close to CBR.
CBR has shown its success in medical diagnoses over hypothetico-deductive reasoning in a comparative study [21], but most medical CBR-systems are oriented to solve classification and diagnostic problems. A few examples exist of case-based medical risk estimation [22,23] but not of epidemiological risks [24], although case-based risk estimation has been widely applied in financial, technological and environmental risk management.

2.2. Constud

Machine learning (ML) and prediction system Constud was created (1) to calculate indices of spatial pattern to a database table or to a binary raster file; (2) to search and weight features and exemplars needed for the most reliable similarity-based predictions; (3) to predict nominal and numerical variables to a database table or to a binary raster file. The system consists of three parts: (1) layers of explanatory variables and pre-classifiers; (2) a knowledge base of observation data (feature vectors), parameters of machine learning, results of machine learning, and predicted values; (3) a software application. The first part of the system is applied only for spatial data. The other components—knowledge base and software—are applicable also for tabular data, as used in this investigation.

The main differences of Constud and traditional methods of statistical modelling are as follows:

1. The estimations are derived from the most similar feature vectors. Generalization in the form of a model is not created.
2. Search of the best solution is a continuous iterative process. Experience obtained during the process is saved in a knowledge base as values of actuality. Features and cases can be added to and excluded from the knowledge base without interrupting the learning process.
3. Feature vectors, parameters and results of ML and metadata are kept in a single knowledge base.
4. The hitherto best set of weights for features and exemplars can always be used for predictions.

The software code of Constud was written in Microsoft Visual Basic NET by Kalle Remm. The compiled application is freeware available from the authors.

2.3. Learning

ML in Constud is an iterative search for the best set of weights for features and for exemplars. Not all cases are useful as exemplars—some repeat each other, some are exceptional or out of date, and some are noisy or not reliable enough. The same is true for features. In addition, not all features are relevant for recognition of a predictable variable. When similarity between cases is calculated we must always ask: similarity in which aspect? The combination of weights of selected exemplars and features that enables the most exact estimations of a predictable variable is called the best predictive set.

The number of features and exemplars in a predictive set needs to be as small as possible according to the principle of parsimony (Occam’s razor). In addition, the lower the number of features and exemplars, the less time is required for calculations both during learning and prediction.

2.4. Calculation of similarity

Similarity between cases is calculated initially as a partial similarity of single features. The algorithm of calculation of similarity depends on the type of a particular feature: nominal or numerical. If the categories of a nominal variable match then the partial similarity is equal to 1, if not, then to 0. The system enables the use of matrices that define similarity between nominal classes but this option was not used in this study. In the case of a continuous feature \( f \), the difference \( D \) between its values \( (T_f \text{ and } E_f) \) for an exemplar \( E \) and a training instance \( T \) was calculated as

\[
D = \frac{|T_f - E_f|}{2 w_E w_f} \tag{1}
\]

where \( w_E \) is the weight of exemplar \( E \), \( w_f \) is the weight of feature \( f \). The partial similarity \( S_f \) between an exemplar and a training instance regarding a feature \( f \) received a value \( 1 - D \) if \( D < 1 \); otherwise \( S_f = 0 \). Further details are given in Remm [25]. The total similarity is calculated as a weighted average of partial similarities. Missing values of some features do not hinder the similarity-based estimations in Constud if any feature has values to calculate the total similarity.

2.5. Decision making

If only one exemplar is used in a similarity-based estimation, the predicted value is transferred from the exemplar. In most cases more than one exemplar is used (the \( k \) nearest neighbour’s method—\( k \)-NN). The \( k \)-value (number of exemplars used in estimations) is controlled computationally as the sum of similarity sought for a decision. It is a parameter in Constud whose value is optimized together with the feature weights. Dynamic optimization of the \( k \)-value was used already in MLNN software [25], which
preceded Constud, and is also proposed by Park et al. [8] and Ahn and Kim [26].

The initial value of \( k \) is set by the user. The process of similarity-based estimations in Constud starts from a similarity level of 100%. If the total similarity of exemplars at this similarity level is less than the sum of similarity sought for the decision, the level is lowered by 1%. The search for the most similar exemplars continues until the necessary sum of similarity is obtained or there are no more similar exemplars. Decision making follows after the search is completed. The predictive system recognizes better the more numerous categories and is less sensitive to noisy data if the sought similarity is large, and is more sensitive to both smaller units and also to errors if the required similarity is less.

The initial value of necessary total similarity in this study was set to 3. Presence/absence of enterobiasis was estimated for each child in the ML of the risk of enterobiasis. If a person was more similar to the selected cases of enterobiasis compared to healthy cases in regard to features selected as relevant, the person was classified as having a high risk of enterobiasis. Persons more similar to the healthy cases received a low risk estimation.

2.6. Objective function

Goodness-of-fit of the ML predictions is estimated by leave-one-out cross validation (LOOC) in Constud. LOOC means that the predicted value for every case is calculated by all exemplars, leaving this case out. The fit of a binomial variable is calculated in Constud as a modified true skill statistic (TSS): the share of true positive similarity plus the share of true negative similarity minus one [27]. TSS assigns equal weight to both correct negative and correct positive estimations and is similar to the Pearson’s coefficient of correlation having an interval of possible values between \(-1\) and \(+1\). TSS expresses absolute fit between observations and estimations, \(-1\) means that all estimations are false, TSS is equal to 0 if half of the positive and half of the negative cases have been estimated correctly. A high level of fit could also be obtained by ignoring the less common value of a binomial variable, e.g. if positive cases form only 1% of observations and all cases are estimated as negative, the share of correct estimations is 99%, but the equal weight objective function indicates that results are correct only at 50%. Traditional TSS uses the share of true positive cases and the share of true negative cases. The modification is that the similarities to positive examples and similarity to negative examples are summed rather than simply counting true positive and true negative predictions. The summed counts are divided by the total sum of similarity yielding the shares of similarity to the category positive and to the category negative.

2.7. Parts of ML iterations

ML iterations in Constud consist of five parts: (1) selection of features, (2) weighting of features, (3) selection of exemplars, (4) weighting of exemplars, and (5) change of actuality values of features and cases. The selection of cases proceeds by adding features one by one according to their predictive value. All cases in a sample are used and the features are added in order of gradually decreasing value. The set of features that gave the best goodness-of-fit continues in the feature weighting stage. The weights of selected features are increased and decreased by a small random value, if the goodness-of-fit improves, the new weights are applied. Selection of exemplars follows using the best weights found during the previous stage. After that, weights are optimized for exemplars, one by one. The weights of features and exemplars are stored in the knowledge base if a hitherto best fit has been reached.

The ordinal initial value of the actuality of features and cases is 100. The actuality of features and cases that were selected into a random sample and were useful in prediction is increased proportionally to the weight. The actuality of features and cases that were in the sample but were excluded from the best predictive set of the weights reached during the ML iteration is decreased. The actualities affect random sampling of features and cases during ML. The features and cases that have a higher actuality have a greater chance to be included in a random sample. The sensitivity of actualities (credulity/credulity of the system) can be set by the user.

2.8. Procedure of ML iterations

The main types of ML available in Constud are as follows:

1. A single iteration using all features and cases is efficient if the number of features is less than 30 and the number of cases less than 1000.
2. Iterative learning using random samples of features and cases is suitable if the number of features or cases is larger and the user prefers not to intervene in the ML process.
3. Iterative learning using small (a few hundred) samples of cases without feature weighting. The best features obtain large actualities relatively quickly, then the useless features are excluded and the sample size is increased by the user. The
method is preferable if the dataset is large and the user is willing to speed up the process by controlling it (and having some subjective influence on the results).

3. Materials and methods

3.1. Data collection

The investigation of enterobiasis was conducted among nursery school children from four counties in Estonia (Tartu, Põlva, Pärnu, and Järva) from 2002 to 2006. The following four principles were followed to select nursery schools: (1) nursery schools had to be spatially scattered in the county; (2) rural areas, small towns and larger towns (if present in a county) had to be represented; (3) consent from the head of a nursery school and from the parents to conduct the investigation was obligatory; (4) the number of children involved from every county was about 500. All children who were present in the nursery school on the day of investigation were examined, except in Tartu county, where only one parallel group was examined in nursery schools containing many parallel age groups.

The research data were obtained from three sources: (1) anal swabs from children for the diagnosis of enterobiasis; (2) close-ended questionnaires for children’s parents; (3) structured observations in rooms of nursery school groups and structured interviews with supervisors of the groups: teachers and teaching assistants.

Infection of *E. vermicularis* was examined using the anal swab technique. The swabs were taken after breakfast just before the children went outside. The presence or absence of pinworm eggs was determined from swabs by microscopy. The total number of studied children was 1905.

The average number of samples in a group was 12 (range 3–26). Observation of rooms and interviews with supervisors took place in all nursery school groups. At a total of 1593 (83.6%) questionnaires were returned. Data on the children were included in the research even if their parents did not return the questionnaire (Table 1). In this case, we still recorded the child’s gender, presence/absence of the infection, and in most cases were able to determine the age of the child (except in 130 cases).

The aim of the questionnaire was to identify possible factors for infection risk associated with children’s households and homes, since enterobiasis spreads mainly in the indoor environment. The questions concerned six topics (the number of questions is in brackets): (1) personal data (5); (2) household members and pet animals (3); (3) previous occurrence of helminthiases in the family (2); (4) socio-economic status of the household (2), (5) living conditions at home (8); (6) the child’s habits and personal hygiene (3).

The aim of the interviews and the observation of rooms, which took place in every nursery school group, were to identify factors of possible infection risk related to the group. The observation took place only in those rooms where children spent considerable time: in play rooms, eating rooms, bedrooms, and washing rooms. The observation gave information about the number, purpose, sanitary conditions and state of repair of the rooms. Interview questions concerned cleaning of the rooms and children’s hygiene, habits and cleanliness.

3.2. Case-based risk estimation using Constud

The above-mentioned attributes of children and nursery school groups were combined in a Microsoft Access database. The multichoice issues from the interviews and questionnaires, such as the method of cleaning rooms at home, that is, dry cleaning, vacuum cleaning, wet cleaning, or any combination of methods, were broken down into several simple Boolean features such as wet cleaning only (yes/no), vacuum cleaning only (yes/no), combination of wet and vacuum cleaning (yes/no). In all, the number of explanatory features for every child attained 77. Seven features concerned the child’s gender, age, habits and personal hygiene, 15, the household and family, and 55, the nursery school.

Ten copies of the database were prepared for parallel ML processes, because the ML process in Constud is not strictly predetermined, as it involves

<table>
<thead>
<tr>
<th>Table 1 The number of observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Children</td>
</tr>
<tr>
<td>----------</td>
</tr>
<tr>
<td>Tartu county</td>
</tr>
<tr>
<td>Põlva county</td>
</tr>
<tr>
<td>Pärnu county</td>
</tr>
<tr>
<td>Järva county</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>
random change in the search for the best weights for features, and random decisions if several options are equal. ML was conducted with all these databases separately in 600 iterations using random sets of 500 cases and 20 features. Three ML iterations using all training data followed with all 10 databases.

The probabilities of statistical significance of the difference between proportions were calculated in Statsoft Statistica 6.1 using the two-sided test.

### 3.3. Comparative risk estimation

Risk of enterobiasis was modelled also using logit-regression and classification trees in Statistica 7.1 (Statsoft). Only those explanatory variables that occurred in the best predictive sets of the 10 separate ML processes were used (Table 2). In addition, the multi-nominal features were included two at a time in all possible combinations as reliable estimates could not be computed for larger model

#### Table 2 TSS of 10 parallel ML processes, estimated overall prevalence, features selected to the best predictive sets and the mean weight of features (>1 in bold)

<table>
<thead>
<tr>
<th>Features</th>
<th>Estimated overall prevalence</th>
<th>The best sets of features from 10 parallel ML processes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TSS</td>
<td>0.38</td>
</tr>
<tr>
<td>Private</td>
<td></td>
<td>.45</td>
</tr>
<tr>
<td>Age (N)</td>
<td>0.75</td>
<td></td>
</tr>
<tr>
<td>Family</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of children (N)</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>Occupation of other children (M)</td>
<td>1.35</td>
<td></td>
</tr>
<tr>
<td>Other children at school or nursery school</td>
<td>1.07</td>
<td></td>
</tr>
<tr>
<td>Pet animals (M)</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td>Heating system (M)</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>Separate rooms for children (M)</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>Carpets (M)</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td>Cleaning frequency (M)</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td>Child’s habits</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Washing hands before meal (M)</td>
<td>0.71</td>
<td></td>
</tr>
<tr>
<td>Child’s hygiene-related habits (M)</td>
<td>0.34</td>
<td></td>
</tr>
<tr>
<td>Nursery group</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Region (M)</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>Most children clean</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td>Some children not clean</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>Neatness of children (N)</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>Equal-age group</td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td>Mean age of children (N)</td>
<td>1.54</td>
<td></td>
</tr>
<tr>
<td>Mixed-age group</td>
<td>0.49</td>
<td></td>
</tr>
<tr>
<td>No. of rooms (N)</td>
<td>0.44</td>
<td></td>
</tr>
<tr>
<td>Washing conditions good</td>
<td>0.23</td>
<td></td>
</tr>
<tr>
<td>Washing conditions poor</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>Few soft toys</td>
<td>0.49</td>
<td></td>
</tr>
<tr>
<td>No. sof toys</td>
<td>0.17</td>
<td></td>
</tr>
<tr>
<td>Rooms airy and well-lighted</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td>Rooms stuffy</td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td>Rooms clean</td>
<td>0.70</td>
<td></td>
</tr>
<tr>
<td>Rooms a bit dirty</td>
<td>0.19</td>
<td></td>
</tr>
<tr>
<td>Range of children’s age (N)</td>
<td>1.48</td>
<td></td>
</tr>
</tbody>
</table>

Feature type: N, numerical; M, multi-nominal; other features are binomial.
designs. The observations were divided into equal size training and validation samples using random index. Predicted probability was classified to positive/negative at level 0.5 for the calculation of TSS. All variants of logit-models were calibrated separately using the total training sample and a reduced sample that contained equal number of positive and negative cases. The ratio of positive and negative cases in the validation sample remained unchanged. The total number of calibrated models was 72.

The best-fit alternative to case-based predictions was obtained with a classification tree consisting of 40 splits and 41 terminal nodes and using 16 explanatory variables (choices in Statistica 7.1: C&RT search, Goodness-of-fit Chi-square, equal prior probabilities, FACT-style stopping at 0.025).

4. Results

4.1. Goodness-of-fit and the best predictive sets

The highest TSS attained in the 10 ML processes was 0.381; the mean was 0.368. Nearly as good estimations can be obtained from different sets of features. The number of features in the best predictive sets of the 10 processes was ranged between 9 and 17 (mean 12.5) out of 77 possible. The overall best set contained 10 features and 640 exemplars (198 infected and 441 healthy) selected from 1905 observations. The TSS training fit of the best logit-model was < 0.244 in simple training sample and 0.439 in the equalised training sample; the best validation fit remained at 0.237. The best classification tree estimated the risk at TSS = 0.775 in training data, 0.18 in the cross-validation sample and 0.35 in all data.

A total of 847 children were more similar in the features selected by the ML-process to the exemplars of infected children selected by the same ML-process and 1058 were more similar to the healthy ones. The most similar exemplars for the largest number of children (79.3%) were all either exemplars of infected or non-infected children, 20.7% of children were similar to both categories to a certain degree (Fig. 1). Healthy children were predominantly correctly recognized by the CBR-system but about a half of the predicted cases of enterobiasis were not confirmed by the single anal swab laboratory analysis. The overall prevalence was 22.8%. The case-based risk estimations varied between 34.5% and 47.8% (average 42.8%) in the results of 10 parallel estimations.

4.2. Risk estimation at the level of an individual and in various groups

The CBR-system enabled us to compare every child with database records and give similarity-based estimations of the infection risk. Individual level estimations can be summarized to a group level. Estimations for three hypothetical children are given as an example.

Person 1: a single child, has an indoor pet dog at home, does not have a separate room at home, always washes hands before meals, sometimes puts his/her finger into his/her mouth, attends a mixed-age nursery school group in which the average age of children in the group is 2.9 years (age range 4 years), and in which not all children are clean, and rooms are well-ventilated—is predicted by the expert system to be non-infected.

Person 2: has an older brother attending elementary school with whom he does not share a bedroom, has an indoor pet cat at home, does not always wash hands before meals, likes to play with his/her cat, attends a nursery school where the average age of children in the group is 4.6 years (age range 2 years), where most children are clean, and where rooms are not well-ventilated—is predicted to be infected.

Person 3: has a sister in a nursery school, does not have pet animals at home, does not have a separate room at home, always washes hands before meals, has no reported special hygiene-related habits, attends a nursery school where the average age of children in the group is 6.2 years (age range 2 years), where most children are clean, and where rooms are well-ventilated—has 54% similarity to the non-infected exemplars and 46% to the infected ones and is estimated as probably non-infected.

Figure 1 Distribution of similarities between cases and exemplars in estimations.
4.3. Main risk factors

There were three features always selected to the best predictive sets: (1) mean age of children in the nursery school group that the child attends, (2) the range of age in the group, and (3) presence and occupation of other children in the family (Table 2). The other features were mutually replaceable or complementary, e.g. features: no soft toys and few soft toys describe the possible involvement of soft toys in the dissemination of *E. vermicularis*. The co-occurrence of antipode-features, such as rooms airy and well lit and rooms stuffy means assigning double weight to the factor: lightness and ventilation. The weightiest features belonged to four groups: (1) age, (2) communication partners, (3) hygiene-related habits, and (4) cleanness of rooms in the nursery school.

The prevalence was higher among older preschool children, being 12.9% (n = 450) among <4-year-old children, and 26.9% (n = 1325) among older children. The similarity-based risk-estimations yielded an even greater difference, exaggerating the risk in the older group (Fig. 2). The relationship is also evident if the mean age of children in a nursery group is considered. The CBR-estimations of the age-dependent risk were 18.9% and 52.0%.

The prevalence was different in groups consisting of children having little age difference, and in mixed-age nursery school groups. The prevalence was less among children belonging to nursery school groups in which the children’s age range was less than 4 years, and was higher if the age-difference was larger, being respectively 19.9% (n = 1454) and 33.2% (n = 451). CBR-estimations of risk were respectively 36.1% and 71.4%. These differences in shares are significant at \( p < 0.0001 \).

The second risk factor — communication partners — is partially included in the age-related features: mean age and range of age of children in a nursery school group. The mark given to nursery school groups for general cleanliness and neatness of their members appeared to be a good indicator of playmate-related risk. The prevalence and estimated risk in case of different cleanliness marks were as follows: poor: 35.7%, 78.6%, n = 28; satisfactory: 28.9%, 52%, n = 304; good: 21.9%, 45.2%, n = 1196; excellent: 18.0%, 27.5%, n = 200.

The risk-relevant communication partners at home were other children and pets. The prevalence was significantly lower among single children (19.4%, n = 98) and among children having only older siblings (>11 years old) (20.3%, n = 232) than among children from families including more than one <11-year-old child (29.7%, n = 935, \( p < 0.005 \)). No remarkable differences in prevalence related to the occupation of other <11-year-old children (at home, nursery school, and elementary school) were discovered. The similarity-based risk estimations were correspondingly 39.8%, 40.1% and 65.5%.

Similarly to the presence of other children in the family, pets also increased the risk of enterobiasis. Both the observed prevalence and the estimated risk were greater in small and moderate-size families when the family had pets (Table 3).

Washing hands before meals was frequently among the relevant features in the best predictive sets (Table 2), although, as a single factor, its variants (always, often, and sometimes) were not related to remarkably different prevalence and risk estimation. The need to wash hands before meals becomes evident when combined with other factors, e.g. the prevalence among children who are

### Table 3

<table>
<thead>
<tr>
<th>No. of children in family</th>
<th>No. of children</th>
<th>No. of children with pet-animals</th>
<th>Prevalence (%)</th>
<th>Estimated risk (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>425</td>
<td>192</td>
<td>13.5</td>
<td>16.7</td>
</tr>
<tr>
<td>2</td>
<td>708</td>
<td>427</td>
<td>19.6</td>
<td>26.5</td>
</tr>
<tr>
<td>3</td>
<td>292</td>
<td>199</td>
<td>24.7</td>
<td>32.4</td>
</tr>
<tr>
<td>4 or more</td>
<td>166</td>
<td>102</td>
<td>38.9</td>
<td>36.9</td>
</tr>
</tbody>
</table>
cuddle their pets was somewhat enhanced for those children who do not always wash their hands before meals, being 27.7% (n = 177) compared to 21.1% (n = 275) among children who always wash their hands before meals (p = 0.1). The prevalence and the estimated risk were high in all cases if the child had any extra habit, be it a pet for cuddling, finger sucking, nail biting, or putting things into their mouth (24.7%, 51.2%, n = 1248). If extra habits were not reported, the prevalence remained at 19.2% (p < 0.005) and the estimated risk was 31.5% (n = 657, p < 0.0001).

Features describing rooms of a nursery group received relatively low weights or were not present in the best predictive sets. As a result of some parallel ML processes, somewhat higher estimated risk was attributed to nursery school groups in which there were few soft toys. The influence of few soft toys may be a result of toys moving from the hands of one child to another more often than in cases of plentiful toys.

The recorded characteristics: gender, finger sucking, mother’s education, previous occurrence of parasitases in family, apartment/private house, dry/water closet, method and frequency of cleaning rooms at nursery school, washing hands and cleaning toys at nursery school were found to be insignificant, as they were not selected into the best predictive sets.

5. Discussion

There are two major advantages of Constud, in addition to the goodness-of-fit of predictions, compared to the tested regression and classification tree models. First, LOOC-validation does not demand exclusion of a large amount of observations from training data to a validation sample. A large number of explanatory features yields in many nearly equally effective sets of predictive features but does not hinder the calibration of a similarity-based prediction system. Second, observations containing empty fields are not used in the calibration of regression models and in the classification tree module in Statistica 7.1. Constud is unaffected by missing feature values. The similarity between two cases is calculated using only those features that have values.

The larger number of false-positive estimations can be attributed to the objective function, where the shares of true positive and true negative have equal weight, and by the lability of actual infection compared to the more stable risk of infection. The detected infection is even more difficult to predict since not all cases of infection are discovered by anal swab analysis. Although it is difficult to take repeat samples in a mass survey, it is recommended in studies of enterobiasis. Investigations have demonstrated that multiple tests increase the likelihood of discovering infected persons. It has been deduced that the rate obtained by double examination is about 4% higher than from a single examination for an infection level of about 10% [28]. Gale [1] argued that surveys based on a single test tape are likely to underestimate the true prevalence by 30–50%.

Remm and Vipp [29] found that the estimated prevalence increased by 7.4% when using double tests, that is, 31.8% compared to 23.2% estimated by the single swab analysis. Consequently, it is highly probable that children estimated by similarity as infected, but non-infected according to the swab analysis, were actually enterobiasis-positive or have been positive or will be positive in the near future.

The case-based risk estimations enhanced differences, as is generally expected from an indicator— in general the differences in risks were much larger than the differences in prevalence. For instance, the risk was estimated to be five times higher in the groups that got a mark of poor for general cleanliness than in the groups assigned excellent—the prevalence was, however, barely two times higher. The same tendency was observable in age-dependent estimations (Fig. 2).

According to Park et al. [30] the egg-positive rate for boys was significantly higher than that of girls. Pezzani et al. [4], Çeliksöz et al. [31] did not find significant dependence of enterobiasis frequency for gender. The gender of children was also unrelated with the prevalence in our study.

Age-related features were valuable for risk estimations in several modifications: age of the child, the mean age of children in the nursery school group, range of the children’s age in the group. The occupation of the other children in the family was also in part an age-related characteristic, as the occupation depends on the age of the other children. The risk of enterobiasis increases among the older nursery school children up to the age of 7. This relationship has been confirmed in earlier studies [5,2], although some investigators have not found a relationship between the age of children and the prevalence [4,28,30–32].

In previous studies, predominantly the age of each child has been related to the infection risk, not the mean age in their nursery school group. The extent of variability of the age of children in a group has not been analysed until now according to our knowledge. Mixed-age groups with an age range over 4 years turned out to be a significant risk factor for the younger members of the group. The prevalence is usually low among children aged 1–4, in
mixed-age groups they are likely to be infected by the older companions.

Pinworm is a contact-helminth: the infection is distributed by communication partners, who, therefore, deserve special attention. The child’s communication partners in the family most attributable to infection are other children and pets. A distinct category that has a significantly lower risk according to our data was only nursery school children in families without pets. John et al. [3] did not find statistically significant correspondence between the size of the family and the risk of enterobiasis. We found an elevated risk as the number of children in a family increases, or when a pet is a member of the family. In addition, the cleanliness of the other members of the nursery school group as communication partners had a prognostic value in our data.

Washing hands before meals and other personal hygiene-related habits of a child were chosen to the best sets of features by most parallel ML processes in our study. Other investigators have found that playing on the floor, nail biting, failure to wash hands before meals [3] and finger sucking [32] are related to the prevalence.

Standards for nursery school rooms and for cleaning of the rooms are quite strictly prescribed in and enforced by Estonian legislation. Therefore, nursery schools are not overcrowded and the differences in general sanitary conditions and state of repair between nursery schools were not large. Thus, most of the 33 features describing nursery school rooms were not included to the best predictive sets of features. In addition, features describing housing and rooms at home were rarely useful for the estimation of the risk of enterobiasis. In Pezzani et al. [4], contrary to expectations, the risk of getting this helminthosis was unrelated to most of the socio-cultural and environmental factors studied (habitat, consumption water, disposal of waste, garbage collection, bathroom characteristics and overcrowding), although the type of housing categorized as satisfactory/unsatisfactory was related to this parasitosis. This contradictory situation could be due to the fact that people living in poor housing have hygienic practices favouring the distribution of eggs in the atmosphere (e.g., by shaking sheets, using a feather dust for cleaning).

Although several authors have pointed to the significance of living conditions [3,5], characteristics of rooms seem to be less significant risk factors than communication partners. According to Pezzani et al. [4] a statistically highly significant association was found between the presence of enterobiasis and the variable affiliations, where the risk category was "being the son/daughter of". There was a higher frequency of enterobiasis also among mothers, probably due to the dense daily contact with their parasite-infected children. Overcrowding at home, considered by Cazorla et al. [32] and Acosta et al. [33] as risk-relevant, could be considered a characteristic belonging mainly to the general factor communication partners and not to dwelling conditions.

In our data, similarly to the findings of Song et al. [5] and Pezzani et al. [4], the risk of enterobiasis was not related to the socio-economic status of the family. Conversely, in other studies, poverty has been noted as a risk factor [32,33]. Çeliksoz et al. [31] found that the infection rates of *Enterobius vermicularis* in primary schools in Sivas (Turkey) depended on school location. Enterobiasis is more frequent among primary school children in urban slum areas than in urban central areas. These findings support the opinion that poorer regions show higher enterobiasis rates. The relationship between poverty and the risk of enterobiasis probably varies in different countries, as does the meaning of poverty.

**Acknowledgements**

The investigation was supported by the Doctoral School of Ecology and Environmental Science and by the Estonian Ministry of Education (SF0180052s07). The authors express their gratitude to students from Tartu School of Health Care for participating in data collection and laboratory analyses, to three anonymous reviewers for valuable comments on an earlier version of the manuscript, as well to Ilmar Part and Robert Szava-Kovats for linguistic corrections.

The experiments comply with the current law of the country in which they were performed (Estonia).

**References**


