*MEDINFO 2017: Precision Healthcare through Informatics*
*A.V. Gundlapalli et al. (Eds.)*

*doi:10.3233/978-1-61499-830-3-447*

# Diagnostic Machine Learning Models for Acute Abdominal Pain: Towards an e-Learning Tool for Medical Students

**Piyapong Khumrin[a], Anna Ryan[b], Terry Judd[b], Karin Verspoor[a]**

[a] *Dept of Computing and Information Systems, School of Engineering, University of Melbourne, Melbourne, Australia,*
[b] *Dept of Medical Education, Melbourne Medical School, University of Melbourne, Melbourne, Australia*

**Abstract**

*Computer-aided learning systems (e-learning systems) can help medical students gain more experience with diagnostic reasoning and decision making. Within this context, providing feedback that matches students' needs (i.e. personalised feedback) is both critical and challenging. In this paper, we describe the development of a machine learning model to support medical students' diagnostic decisions. Machine learning models were trained on 208 clinical cases presenting with abdominal pain, to predict five diagnoses. We assessed which of these models are likely to be most effective for use in an e-learning tool that allows students to interact with a virtual patient. The broader goal is to utilise these models to generate personalised feedback based on the specific patient information requested by students and their active diagnostic hypotheses.*

***Keywords:***

Decision Support Systems, Clinical; Formative Feedback; Artificial Intelligence

## Introduction

Diagnostic reasoning is the cognitive process of deriving a correct diagnosis from a patient's presenting clinical problem. The development of diagnostic reasoning skills in medical training starts with a disease-oriented approach by learning common presentations of different diseases [1]. Given this knowledge, students approach patients with a presenting problem and hypothesise the most likely diagnosis. They gather patient information through the steps of history taking, physical examination, and consideration of the results of laboratory procedures and other investigations. Students accumulate the information to prune and prioritise possible diagnoses until they get to a final diagnosis [2]. During medical training, students practice diagnostic reasoning skills with patients under expert supervision, called "bedside teaching". Students gather patient information and present their diagnostic reasoning to the expert. The expert identifies errors, misconceptions and inadequacies and formulates suitable feedback to help students to reconstruct their knowledge [3]. Personalised feedback is a key element of learning and instruction [4-6] and bedside teaching is known to improve diagnostic reasoning skills. However, its use is declining due to a range of factors, including increased patient turnover, concerns with patient privacy, increased technology in the diagnostic process, increased numbers of students, and limited availability of experts [7; 8].

E-learning systems can help to address some of these challenges and when used effectively, in conjunction with traditional approaches, can aid in the development of diagnostic reasoning skills [9-11]. However, tailoring the feedback from e-learning systems so that it is both effective and fits the needs of individual students is difficult [4; 12]. Effective feedback should help the student to identify what they already know or have mastered, where potential knowledge gaps or misconceptions lie, provide an indication of their learning progress, and support them to achieve their learning goals [13].

Clinical Decision Support Systems (CDSS) are computer systems that assist doctors to make decisions and are typically used in either the diagnostic process or to support clinical management. CDSS synthesise information based on patient data and use the information to generate a prediction. Prior research has demonstrated that using a CDSS to assist physicians' diagnostic and treatment processes can improve both the effectiveness and efficiency of patient care [14-16], a clear example being the management of acute abdominal pain [17]. Applying the concept of a CDSS, we aim to use machine learning models to produce personalised feedback within an adaptive game-based learning tool intended to support the development of medical students' diagnostic reasoning and decision-making skills. The learning tool will allow students to interact with a virtual patient, and revealing relevant patient information and diagnostic predictions in response to the students' actions and requests.

This paper presents the development of the machine learning model that underpins the e-learning tool. It describes the collection and processing of a large corpus of patient data, and the development, training, testing and comparison of alternative machine learning models based on these data. A preferred model is identified and the rationale for its selection in the context of the e-learning tool explained. We also discuss how this model may be leveraged within the tool to generate personalised and appropriate feedback for medical students.

## Methods

### Phase I: Data collection

Electronic patient records were collected to train the machine learning models. We captured and processed de-identified medical data from three disparate but complementary sources; 1) Student log cases 2) Student entered Electronic Health Records (created by medical students at our university as part of their course and stored within their curriculum delivery system 3) Electronic health records within the public hospital affiliated with our university's medical school. The data were used to develop clinical scenarios to be presented within the learning tool and to train machine learning models for of the learning tool's embedded CDSS. All data collection was approved either by our university's Health Sciences Human Ethics Sub-Committee or by the affiliated hospital's Human Research Ethics and Research Committee.

### Phase II: Case selection

We selected cases from the electronic systems with the following inclusion and exclusion criteria:

***Inclusion criteria***

1. A principal diagnosis of one of five key conditions (appendicitis, gastroenteritis, urinary tract infection, ectopic pregnancy, or pelvic inflammatory disease).
2. The treatment protocol for the principal diagnosis was completed.

***Exclusion criteria***

1. More than one of the conditions of interest were diagnosed in an individual patient in the same admission.
2. A previous history of other conditions or procedures or treatments that would rule out one or more of the key diagnoses (e.g. appendectomy for appendicitis).

Two hundred and ninety-eight clinical cases were identified by applying the query criteria. Ninety cases were excluded leaving 208 valid clinical cases (see Table 1).

### Phase III: Data pre-processing

PK manually extracted data from the selected cases and transformed the information into an array of features. Seventy-five features were extracted from history taking (n=48), physical examinations (n=13), laboratory and investigation results (n=13), and a target class (n=1).

*Table 1 – Sample size per diagnosis*

| Diagnoses | n |
|---|---|
| Appendicitis (AP) | 51 |
| Gastroenteritis (GE) | 53 |
| Urinary tract infection (UTI) | 68 |
| Ectopic pregnancy (EP) | 11 |
| Pelvic inflammatory disease (PID) | 25 |
| Total | 208 |

### Phase IV: Machine learning training

We used pre-processed clinical features from Phase III to train machine learning models for classifying the features into one of the five target diseases. We utilised Weka version 3.8 [18] for the training process using 10-fold cross validation on a training set to evaluate alternative algorithms: testing Naïve Bayes, Support Vector Machine (SVM), Neural Networks (NN), C4.5 decision tree (J48), and Logitboost (using DecisionStump as a classifier). We used the correlation attribute evaluation to rank the level of feature relevance to predict a diagnosis. We intend to use one or more of the machine learning models to predict diseases when students raise possible diagnoses on the basis of clinical observations. We plan to transform the prediction of the classifier to a suitable form of feedback to represent the likelihood of diagnosis based on present findings. The rank of a given feature's relevance in the model will be used to guide students' feature selection.

We grouped the target classes in two ways. First, we treated the entire dataset as a single group and targeted classification of the five diseases considering all diagnoses together, which is a "multi-class classification" scenario. Second, we divided the target classes into two groups, to create a "binary classification". In this case, the first group contains one of the target diseases, with all others being assigned to the second group. The diagnoses in the second group are then merged into a single class – e.g. the classifier makes a decision between "appendicitis" and "not appendicitis".

## Results

### Multi-class classification result

Table 2 shows the percentage of overall correctly classified instances from six machine learning models. ZeroR (majority class classifier) provides a baseline performance. All classifiers predict better than the baseline but the top three classifiers were Logitboost, NaïveBayes, and Neural Network, shown in bold.

*Table 2 – Accuracy of multi-class classification*

| Classifiers | % Accuracy |
|---|---|
| ZeroR | 32.69 |
| J48 | 66.35 |
| SVM | 72.60 |
| NN | **84.62** |
| NaïveBayes | **85.10** |
| Logitboost | **94.71** |

Table 3 shows the F1-measure of classification in different diagnoses. Logitboost predicts all diagnoses with the highest performance (F1-measure between 90 – 99%), and performed substantially better than all other classifiers in the case of EP.

*Table 3 – F1-measure*

| | AP | GE | UTI | EP | PID |
|---|---|---|---|---|---|
| ZeroR | 0.000 | 0.000 | 0.493 | 0.000 | 0.000 |
| NaïveBayes | 0.887 | 0.857 | 0.889 | 0.571 | 0.793 |
| SVM | 0.755 | 0.755 | 0.786 | 0.455 | 0.480 |
| NN | 0.857 | 0.862 | 0.949 | 0.375 | 0.679 |
| J48 | 0.706 | 0.627 | 0.744 | 0.167 | 0.500 |
| Logitboost | **0.923** | **0.925** | **0.993** | **0.900** | **0.939** |

The distribution of correct and incorrect classifications (confusion matrix) in different diagnoses appears in Table 4 for Naïve Bayes and Table 5 for Logitboost. Diseases in rows and columns represent the true and predicted diagnoses, respectively. AP, EP, and PID have the most instances of misclassification.

*Table 4 – NaïveBayes confusion matrix*

| predicted<br>true | AP | GE | UTI | EP | PID |
|---|---|---|---|---|---|
| **AP** | 47 | 1 | 2 | 1 | 0 |
| **GE** | 4 | 45 | 0 | 0 | 4 |
| **UTI** | 1 | 6 | 56 | 2 | 3 |
| **EP** | 2 | 0 | 0 | 6 | 3 |
| **PID** | 1 | 0 | 0 | 1 | 23 |

*Table 5 – Logitboost confusion matrix*

| predicted<br>true | AP | GE | UTI | EP | PID |
|---|---|---|---|---|---|
| **AP** | 48 | 3 | 0 | 0 | 0 |
| **GE** | 3 | 49 | 1 | 0 | 0 |
| **UTI** | 0 | 0 | 68 | 0 | 0 |
| **EP** | 1 | 0 | 0 | 9 | 1 |
| **PID** | 1 | 1 | 0 | 0 | 23 |

Table 6 shows selected key decision features from J48 decision tree (not shown) which are correlated to clinical knowledge.

*Table 6 – Key decisions for the diagnoses on J48 decision tree*

| Diagnoses | Features |
|---|---|
| AP | Right lower abdominal pain |
| GE | Upper abdominal pain<br>Diarrhea |
| UTI | Dysuria<br>Lower abdominal pain |
| EP | Serum hCG |
| PID | Serum hCG |

### Binary classification result

Table 7 shows the F1-measure score of binary classifications for individual diagnoses, using the NaïveBayes and LogitBoost classifiers. The first and second sub-columns under the classifier column represent the first and second groups in the binary classification, respectively. The last sub-column is the average F1-measure in the first two sub-columns. LogitBoost classifies all target diagnoses with good accuracy whereas NaïveBayes registers a significant drop for EP.

*Table 7 – F1-measure of the binary classifications*

| | Naïve Bayes | | | LogitBoost | | |
|---|---|---|---|---|---|---|
| | X | not X | Avg | X | not X | Avg |
| AP | **0.851** | 0.952 | 0.928 | 0.832 | 0.949 | 0.918 |
| GE | 0.874 | 0.958 | 0.937 | **0.882** | 0.962 | 0.942 |
| UTI | 0.894 | 0.951 | 0.932 | **0.963** | 0.982 | 0.976 |
| EP | 0.500 | 0.969 | 0.945 | **0.952** | 0.997 | 0.995 |
| PID | 0.724 | 0.955 | 0.928 | **0.939** | 0.992 | 0.985 |

### Feature selection

#### *Feature selection on multi-class classification*

The most highly ranked features with respect to the multi-class classification were a history of left abdominal pain and the patient's age. The top ten features, all of which returned a percentage relevance of at least 15% are listed in Table 8.

*Table 8 – Top 10 features of the multi-class classification*

| Ranked features | % relevance |
|---|---|
| Left abdominal pain history | 31 |
| Age | 25 |
| Lower abdominal tenderness | 20 |
| Guarding | 19 |
| Upper abdominal pain history | 18 |
| Vomiting | 16 |
| Rovsing's sign | 16 |
| Rebound tenderness | 16 |
| Leukocyte in UA | 15 |
| Pain quality | 15 |

#### *Feature selection on binary classification*

We also measured the correlation of features to individual diagnoses, based on the binary classification model. Table 9 shows selected features across the five diseases. Symptoms such as age, gender, location of abdominal pain, characteristics of pain, fever, gastrointestinal, urinary tract system, and gynaecological histories are useful in distinguishing between the different diseases. Right lower abdominal pain and pain migration are more specific to AP. Diarrhea is more specific to GE. Urinary tract symptoms are more specific to UTI. UPT and serum hCG are most specific to EP, while gynaecological symptoms and laboratory result for sexual transmitted diseases are most strongly correlated with PID.

*Table 9 – Selected features of the binary classifications*

| Features | Relevant features |
|---|---|
| AP | Right lower abdominal pain, pain migration |
| GE | Diarrhea |
| UTI | Dysuria, urine colour, haematuria |
| EP | UPT, serum HCG |
| PID | Vaginal discharge, PCR for Chlamydia |

## Discussion

### Retrospective vs prospective data collections:

We used a retrospective data collection method to collect clinical cases for training machine learning models. Unlike prospective approaches, this meant we were unable to constrain the records within a single template. Further limitations of this method included the variable formatting and structure of records, missing values, and temporal data. Our reliance on different data sources (even within the one hospital it is not unusual for different clinical departments to use different electronic health record systems) created a range of data entry and processing issues related to feature definition, type and sequence as well as clinical interpretation. This variation introduced considerable noise, reducing data validity and complicating data pre-processing. Many of these issues would undoubtedly have been reduced in the case of prospective data collection, as requirements or recommendations around record creation can be more closely defined and monitored. However, prospective data collection would have required considerable additional planning on our part, and carried substantial additional time and administrative costs that are likely to have been unacceptable to hospital staff given their heavy workloads.

Table 10 shows pros and cons of retrospective and prospective data collection. In summary, retrospective data collection is simpler with lower costs and does not impact on patients' treatment. Prospective data collection, on the other hand, provides higher quality data and data validity but is likely to involve unacceptable costs.

*Table 10 – Pros and cons between retrospective and prospective data collection*

| Factors | Retrospective | Prospective |
|---|---|---|
| Budget | **less** | more |
| Increase work for treatment process | **no** | yes |
| May influence treatment process | **no** | yes |
| Sample size | **flexible** | restricted |
| Quality of data | poor to good | **very good** |
| Data validity | poor to good | **very good** |

For this study, we utilised case records provided in three different digital formats: plain text, scanned documents (images), and Word documents. The majority of these were extracted from the hospital's record databases. The number of records available to us were sufficient for our purposes except in the cases of EP and PID, because patients diagnosed with those two diseases tend to be admitted to a specialist women's hospital rather than the general hospital involved in our study. Inclusion and exclusion criteria were used to filter the cases because certain histories have a critical impact on formulating a list of possible diagnoses. For example, a patient who has had their appendix or ovaries removed should never be diagnosed with AP or EP, respectively. Those histories provide a spot conclusion rather than enhancing the development of diagnostic reasoning processes.

### Data pre-processing

Before PK started extracting data, he listed common presentations of the five diseases from medical standard textbooks [19-23]. He used standard medical terms from UMLS [24] and SNOMED CT [25] to identify and organise features and synonyms. Seventy-two (72) of 75 available features contained some missing values (only age, gender, and target diagnosis were complete). Missing values commonly occur because admitting doctors deem it as self-evident, unnecessary or irrelevant. For example, doctors will never ask questions about menstrual history, or order a urine pregnancy test and serum hCG for male patients. Similarly, a normal urine finding typically infers a negative finding in relation to all abnormalities of the urinary tract.

Symptoms also typically change over time. For example, the location, severity and quality of pain, and whether the pain is relieved by medication, can all change as a disease progresses. In the case of the location of pain, we divided pain location into two episodes – prior to and at admission. If these were different then the pain had migrated. More generally, we only used symptoms and signs from records created in the emergency department or on first admission to a hospital ward to reduce the effects of timing and treatment.

### Clinical interpretation of classification results

The top three classifiers were Logitboost, NaïveBayes, and Neural Network, which all had a predictive accuracy above 80% (see Table 2). The Logitboost algorithm improves the results of a classification by reweighting mis-classified samples and taking a weighted majority vote to form training data [26]. It returned a significantly higher overall predictive accuracy than either NaïveBayes or Neural Network classifiers due to its superior predictive performance for EP (see Table 3). The classification performance for EP was low for all other algorithms due to a combination of low record numbers and missing data. We believe that Logitboost performed well despite these issues because of its use of a Decision Stump; a subroutine within Logitboost that analyses patterns of missing data to develop rules for classification.

When we considered the distribution of misclassified EP over other diseases predicted by NaïveBayes (Table 4) and Logitboost (Table 5), the models were more likely to misclassify EP as either AP and PID, which have a number of overlapping symptoms, rather than UTI or GE. By way of comparison, in a series of early studies, de Domal and colleagues [27; 28] used a CDSS employing a Bayesian classifier to predict abdominal pain within 600 prospectively collected clinical cases. They reported an overall diagnostic accuracy of 91.8%.

Table 11 compares the research methods and classification results for the current and these earlier studies when using a similar classification approach [27]. In this case, the machine learning model developed by de Dombal had the higher overall classification accuracy. We noted the key success factors in the de Dombal study were sufficient samples, equal distribution of sample, and the quality of input data.

While all classifiers apart from the baseline (ZeroR) provided more accurate predictions than the J48 algorithm, it is the only one that produces human-readable output. Its decision tree style output is easily interpreted and has the potential to provide useful feedback to users during the decision-making process.

*Table 11 – Comparison of research methods and classification result between two studies*

| Methods/Result | this study | de Dombal |
|---|---|---|
| Data collection | retrospective | prospective |
| Sample size | 208 | 600 |
| Machine learning method | NaïveBayes | NaïveBayes |
| Target classes | 5 | 7 |
| Overall accuracy | 85.10 | 91.80 |

### Key features

For multi-class classification, left abdominal pain appears to be a key feature as it positively selects for those diseases presenting with lower abdominal pain – appendicitis, ectopic pregnancy, and pelvic inflammatory disease. Gastroenteritis and urinary tract infection are less likely to be selected because the pain position for these two diseases is more diffuse. However, a history of left abdominal pain reduces the probability of appendicitis and indicates more strongly ectopic pregnancy or pelvic inflammatory disease. Age is also an important factor, with appendicitis, ectopic pregnancy and pelvic inflammatory disease most strongly associated with younger patients. In the case of the binary classification, we were able to identify key diagnostic features associated with each disease. We will use this information to help medical students to identify strongly relevant clinical information to support their diagnostic decisions.

### Machine learning model selection

Our final choice of machine learning model for use in the proposed learning tool is informed by two requirements: clinically appropriate predictions; and classification performance. We give precedence to clinical interpretation because proper development of the diagnostic reasoning process is more important than maximising the number of correct decisions. Binary classification is preferred because it provides a better sense of scaling of the likelihood level than multi-class classification. And, while the overall prediction performance of NaïveBayes is slightly inferior to Logitboost, the NaïveBayes predictions are more reflective of actual clinical judgements. For example, Logitboost predicts zero probability of ectopic pregnancy on a female patient presenting with right lower abdominal pain, which is clinically inappropriate. Accordingly, we selected the NaïveBayes classifier in combination with binary classification as our preferred model.

### Feedback representation

We plan to use the machine learning model to provide two types of feedback (interim and final) within the proposed learning tool. Interim feedback will be based on the interpretation of predictive correlations between selected patient information and top three most likely diagnoses following the history taking, physical examination, and laboratory and investigations steps. Final feedback will present the correct diagnosis, and generate a user score based on how often the correct diagnosis is selected in the differential diagnosis list, correlation to the correct diagnosis, and the inclusion of key patient information.

We plan to use an overlap model to assess students' performance (and potential learning gains) while using the tool [29]. This treats the student's decisions (student model) as an incomplete model which can be compared to the machine learning model (complete model). The more similar these two models are, the higher the assessment of the student's performance.

## Conclusion

Timely and clinically appropriate personalised feedback is key to the development of students' diagnostic reasoning skills. E-learning has a role to play here, through the provision of personalised and appropriately scaffolded feedback on students' diagnostic decision-making on virtual patient cases. We propose to use diagnostic models derived through machine learning as the basis for giving relevant feedback. In this paper, we describe how we trained machine learning models using a large corpus of real clinical cases to develop differential diagnoses related to presentations of abdominal pain. We selected a model that combines a Naïve Bayes classifier with binary classification for further use in the learning tool based on a combination of its predictive performance and the clinical relevance of that model's predictions.

## Acknowledgements

PK thanks Chiang Mai University for a scholarship to support his PhD study. We also thank the Tilley family, Dr Chris Leung, and Professor Richard O'Brien of Austin Hospital Clinical School, the University of Melbourne for research support.

## References

[1] R. Patel, J. Sandars, and S. Carr, Clinical diagnostic decision-making in real life contexts: A trans-theoretical approach for teaching: AMEE Guide No. 95, *Med Teach* **37** (2015), 211-227.

[2] A. Venot, A. Burgun, and C. Quantin, *Medical informatics, e-Health : fundamentals and applications*, 2014.

[3] M. Garout, A. Nuqali, A. Alhazmi, and H. Almoallim, Bedside teaching: an underutilized tool in medical education, *Int J Med Educ* **7** (2016), 261-262.

[4] M. Frize and C. Frasson, Decision-support and intelligent tutoring systems in medical education, *Clin Invest Med* **23** (2000), 266-269.

[5] W.C. McGaghie, S.B. Issenberg, E.R. Petrusa, and R.J. Scalese, A critical review of simulation-based medical education research: 2003-2009, *Med Educ* **44** (2010), 50-63.

[6] D.H. Jonassen, Assoc. Educational Communications and Technology, *Handbook of research for educational communications and technology: a project of the Association for Educational Communications and Technology*, Macmillan Library Reference USA, New York, 1996.

[7] M. Peters, O. Ten Cate, Bedside teaching in medical education: a lit review, *Perspect Med Educ* **3** (2014), 76-88.

[8] A. Salam, H.H. Siraj, N. Mohamad, S. Das, Y. Rabeya, Bedside teaching in undergraduate medical education: issues, strategies, and new models for better preparation of new generation doctors, *Iran J Med Sci* **36** (2011), 1-6.

[9] P. Devitt and E. Palmer, Computers in medical education 1: evaluation of a problem-orientated learning package, *Aust N Z J Surg* **68** (1998), 284-287.

[10] J.N. Hudson, Computer-aided learning in the real world of medical education: does the quality of interaction with the computer affect student learning?, *Med Educ* **38** (2004), 887-895.

[11] J.G. Ruiz, M.J. Mintzer, R.M. Leipzig, The impact of E-learning in medical education, *Acad Med* **81** (2006), 207-212.

[12] E. Vasilyeva, Towards personalized feedback in educational computer games for children, in: *Proceedings of the sixth conference on IASTED International Conference Web-Based Education - Volume 2*, ACTA Press, Chamonix, France, 2007, pp. 597-602.

[13] B.J. Mason, R. Bruning, Providing feedback in Computer-Based Instruction: What the research tells us, 2001.

[14] K. Kawamoto, C.A. Houlihan, E.A. Balas, and D.F. Lobach, Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success, *BMJ* **330** (2005), 765.

[15] A.X. Garg, N.K. Adhikari, H. McDonald, M.P. Rosas-Arellano, P.J. Devereaux, J. Beyene, J. Sam, and R.B. Haynes, Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review, *JAMA* **293** (2005), 1223-1238.

[16] R. Kaushal, K.G. Shojania, and D.W. Bates, Effects of computerized physician order entry and clinical decision support systems on medication safety: a systematic review, *Arch Intern Med* **163** (2003), 1409-1416.

[17] J.G. Cooper, R.M. West, S.E. Clamp, and T.B. Hassan, Does computer-aided clinical decision support improve the management of acute abdominal pain? A systematic review, *Emerg Med J* **28** (2011), 553-557.

[18] E. Frank, I.H. Witten, and M.A. Hall, *Data mining : practical machine learning tools and techniques*, Morgan Kaufmann, Burlington, Mass., 2016.

[19] N.J. Talley, S. O'Connor, A summary of the gastrointestinal examination and extending the gastrointestinal examination, in: *Clinical examination: a systematic guide to physical diagnosis*, CL Elsevier, Sydney, 2010.

[20] N.J. Talley and S. O'Connor, Correlation of physical signs and gastrointestinal disease, in: *Clinical examination : a systematic guide to physical diagnosis*, Churchill Livingstone Elsevier,, Sydney, 2010.

[21] N.J. Talley, S. O'Connor, The gastrointestinal history, in: *Clinical examination: a systematic guide to physical diagnosis*, CL Elsevier, Sydney, 2010.

[22] N.J. Talley, S. O'Connor, The gastrointestinal physical examination, in: *Clinical examination: a systematic guide to physical diagnosis*, CL Elsevier, Sydney, 2010.

[23] J. Hall, Essentials of clinical examination handbook, in, Thieme,, New York, 2013, p. 640 p.

[24] D.A. Lindberg, B.L. Humphreys, and A.T. McCray, The Unified Medical Language System, *Methods Inf Med* **32** (1993), 281-291.

[25] K. Bardadin, [SNOMed--international system for coding changes in medicine], *Pol J Pathol* **46** (1995), 203-205.

[26] J. Friedman, T. Hastie, R. Tibshirani, Additive logistic regression: a statistical view of boosting, (2000), 337-407.

[27] J.R. Staniland, J. Ditchburn, and F.T. De Dombal, Clinical presentation of acute abdomen: study of 600 patients, *Br Med J* **3** (1972), 393-398.

[28] J.C. Horrocks, A.P. McCann, J.R. Staniland, D.J. Leaper, and F.T. De Dombal, Computer-aided diagnosis: description of an adaptable system, and operational experience with 2,034 cases, *Br Med J* **2** (1972), 5-9.

[29] J.L. Stansfield, *Wumpus Advisor I. A First Implementation of a Program That Tutors Logical and Probabilistic Reasoning Skills.* Distributed by ERIC Clearinghouse, [Washington, D.C.], 1976.

### Address for correspondence

Professor Karin Verspoor, karin.verspoor@unimelb.edu.au
The University of Melbourne, Melbourne VIC 3010 Australia.