MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press

A novel diagnostic aid (ISABEL): development and preliminary evaluation of clinical performance

Padmanabhan Ramnarayan[a], Amanda Tomlinson[a], Gautam Kulkarni[a], Anupama Rao[b], Joseph Britto[a]

[a]*Department of Paediatric Intensive Care, St Mary's Hospital, London, UK*
[b] *Department of Haematology, Watford General Hospital, Watford, UK*

Abstract

Clinical diagnostic aids are relatively scarce, and are seldom used in routine clinical practice, even though the burden of diagnostic error may have serious adverse consequences. This may be due to difficulties in creating, maintaining and even using such expert systems. The current article describes a novel approach to the problem, where established medical content is used as the knowledge base for a pediatric diagnostic reminder tool called ISABEL. The inference engine utilizes advanced textual pattern-recognition algorithms to extract key concepts from textual description of diagnoses, and generates a list of diagnostic suggestions in response to clinical features entered in free text. Development was an iterative process, relying on sequential evaluation of clinical performance to provide the basis for improvement. The usage of the system over the past 2 years, as well as results of preliminary clinical performance evaluation are presented. These results are encouraging. The ISABEL model may be extended to cover other domains, including adult medicine.

Keywords:

Decision support; diagnosis; evaluation studies; pattern recognition; diagnostic error

Introduction

Biomedical knowledge has grown exponentially in the past few years, resulting in severe information overload for clinicians [1]; it is estimated that this problem will double every 20 years [2]. However, rapid growth has not affected all domains of medical knowledge equally: information related to newer medical tests and treatments is constantly evolving, whereas traditional knowledge pertaining to clinical diagnosis has changed relatively little. Recent techniques that attempt to summarize latest treatment recommendations in line with changing medical evidence are now available, and are popular with clinicians [3][4], whereas systems that might assist in routine clinical diagnosis remain scarce.

We know, however, that medical information relevant to making clinical diagnoses is constantly needed [5], and that this information need is fulfilled mainly by consulting textbooks [6]. We also know that errors related to misdiagnoses, or missed diagnoses, constitute a significant proportion of the preventable burden of medical error [7]. Some diagnostic errors may be due to 'errors of omission' (failure to consider all clinically relevant diagnoses during initial workup). In addition, incorrect formulation of the clinical problem, as well as difficulty in extracting relevant information from textbooks quickly, whether paper-based or electronic, may contribute to diagnostic error.

One reason why computerized diagnostic aids are scarce may be related to the difficulty of converting traditional medical knowledge into computer-readable form. Existing aids for internal medicine, such as DXplain, QMR and ILIAD [8][9][10], were developed over many years, involving the input of multiple experts to provide semi-probabilistic relationships between thousands of clinical features and hundreds of diseases. These tools were also developed to assist the clinician primarily during the rare entity of a diagnostic dilemma (clinical dead-end), by acting as 'oracles'. As a result of this design, they often required the user to expend a considerable amount of time interrogating the system [11]. In order to regularly use such a stand-alone system in practice, a clinician had to be highly motivated, one reason why diagnostic decision support as a concept may not have captured clinicians' interest.

This paper aims to describe the development of a novel diagnostic reminder tool called ISABEL (www.isabel.org.uk) [12][13][14], which utilizes unstructured information from standard medical content, including textbooks to provide a set of diagnostic reminders for any clinical scenario, in response to clinical features entered in free text. In stand-alone format, it is intended to be used by clinicians in routine practice, in a negligible amount of time. An analysis of the system's performance is also described.

Materials and Methods

Underlying knowledge base

Electronic text was used to populate a pre-designed diagnostic tree comprising around 3500 diagnoses. This tree was based on the table of contents derived from one standard textbook (Nelson's Textbook of Pediatrics, 16th Edition). Figure 1 depicts part of this diagnostic tree.

Electronic text pertaining to each diagnosis within the tree was copied into the ISABEL database, without any modifications, by one research nurse. Thus, unformatted text relating to the same disease, from different sources, was collated under one disease label. Where new disease labels were necessary to accommodate

text from a new source, they were created within the same overall diagnostic tree model. Since there was no modification of the text involved, this entire process was easily done.

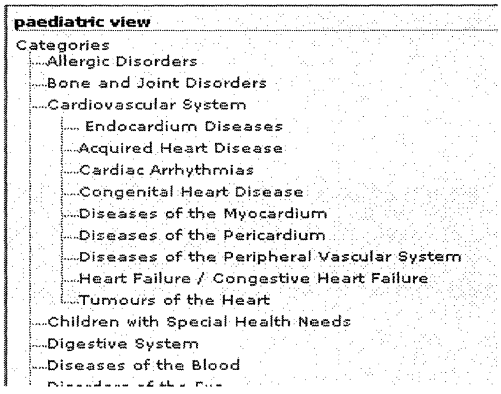

Figure 1 - View of part of the ISABEL diagnostic tree

Inference engine

Commercially available software that employs advanced pattern-recognition techniques on unstructured text to extract a document's digital essence, identifies and encodes the unique signature of key concepts within a document, and creates concept agents to match document profiles with similar ideas as the input text was used. This software utilizes Bayesian Inference and Shannon's principles of information theory to generate its pattern-matching algorithms to enable sophisticated concept extraction from documents.

By aggregating text related to one specific diagnosis under a single diagnostic label within the diagnostic tree, it was possible for the software to generate a unique signature of key concepts for each diagnosis, by using its concept extraction techniques. This signature was constantly modified with the addition of text from each additional source used to populate the ISABEL database.

Search methodology

In response to a set of key clinical features for a patient (concepts), ISABEL utilized the inference engine to search the underlying database of text, and return all documents (diagnostic labels) whose concept signature matched that generated from the clinical features. Clinical features could be altered (by entering additional findings or deleting findings) to reflect a different concept signature, and thus alter the results of the search. Due to the nature of the search mechanism, only clinical features described in textual language were used to generate a concept signature (i.e numerical values, such as the age of the child, were not used).

System architecture and delivery platform

In order to maximize the use of the system, and eliminate inequalities of regional distribution, ISABEL was delivered on the World Wide Web to all medical practitioners, after a short registration process. To this end, a website was created by DynamicWeb, UK using Javascript. The Dynamic Reasoning Engine™ (DRE) was hosted on a dedicated ISABEL server, as was the database comprising the diagnostic tree, which could be construed as being the equivalent of multiple documents of text in html.

At the front end, a free text box, into which the user could enter the clinical features of a patient, was created on a dedicated diagnostic tool webpage on the ISABEL website. On searching the database with these features, a list of all matching diagnostic labels (diseases) was returned. The maximum and minimum number of the diagnoses displayed on the results page could be varied by the developers of the system.

Remodeling the search mechanism

Preliminary examination of the system's raw results by the medical team (consisting of three pediatricians) suggested four problems, which led to a remodeling of the underlying architecture of the ISABEL database.

- Since numerical values were not used in the generation of the concept signature, it became obvious that users had to specify the patient's age group separate from the clinical features to avoid age-inappropriate diagnoses being displayed (such as neonatal meningitis for a 3 year old child)
- Due to the global audience, it was essential to take into account where the patient originated from, and to tailor diagnostic suggestions accordingly.
- Due to the free text (variable) nature of input, it was necessary to create an intermediate filter between the input and the DRE, whose main function was to convert non-medical terminology into medical terms.
- Since the primary function of the system was to provide diagnostic reminders, each holding equal clinical value, it was felt that the degree of concept matching should not be used as the basis for the ordering of the diagnostic suggestions.

These changes were achieved by tagging each of the diagnostic labels in the tree to specific age groups (newborn, infant, child and adolescent), and to specific regions of the world (e.g. North America, Western Europe etc., as per World Health Organization guidelines). In addition, the intermediate filter to convert common non-medical terms into appropriate medical terms was developed specifically for ISABEL by the medical team in 4 weeks. Table 1 shows some common terms included in this filter.

Table 1: Examples of terms included in the filter

Lay terms & abbreviations	Medical translation
Nad	Normal
Hot, high temperature	Fever, pyrexia
WBC	White cell
Shut down, cold peripheries	Shock, vasoconstric-

Additional drop-down boxes were provided for the user to specify the age-group and the region, in addition to the existing free text box for clinical features. Diagnostic results were arranged into body systems to which they pertained (Asthma - Respiratory disorder), rather than in rank order of the degree of concept match. This facilitated a patho-physiological approach to the diagnostic process. Further information regarding each diagnosis

in the suggestion list could be obtained by clicking on it – text from Nelson's textbook was provided for reference. Figures 2 & 3 show how a set of clinical features entered in free text into the search box produce a set of diagnostic labels (with the preceding text: have you considered?) for the user's attention.

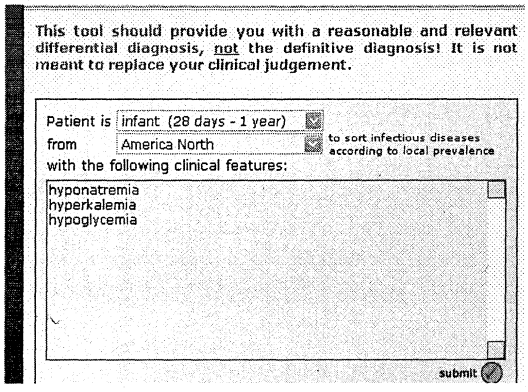

Figure 2 - Clinical features entered into free textbox

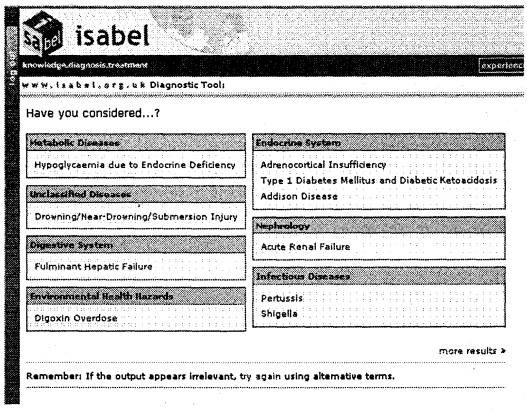

Figure 3 - Diagnostic suggestions arranged to reflect pathophysiological process involved

Measuring usage of the system on the Web

The complete system as described above was available on the Internet from June 2001. Usage statistics were used as indicators of the popularity of the system, and were measured using analog 5.22. Data from July 2001 to date are provided in the results section.

Preliminary estimates of the system's performance

Since the primary role of the system was to offer relevant significant diagnostic reminders for a variety of clinical scenarios in pediatrics in a negligible amount of time, preliminary testing of the performance of the system was based on examining if important diagnostic suggestions were offered, and how long it took to obtain results from the system. This testing was done by the developers of the system, rather than users, in a laboratory setting removed from clinical practice.

Clinical data from 100 real patients, drawn from an unselected consecutive sample of children attending 4 emergency departments in the UK was used. These data consisted of age-group, initial clinical features (including results of available 'first-pass' tests) and final discharge diagnoses. They were collected by clinicians working in these departments for the study, and were not modified by the developers in any way during testing. Cases were examined by a panel of two pediatricians, working together, who produced a bare minimum list of 'significant' diagnoses that ought to have formed part of the examining physician's diagnostic work-up list to ensure clinical safety (gold standard). These clinical data were also entered into ISABEL by one research nurse, and the resultant diagnostic suggestion list was compared to the gold standard list. The maximum number of suggestions was fixed at 15 for this study.

Outcomes: Comprehensiveness ratio: mean value of match between the gold standard list and ISABEL's list (expressed as a proportion).

Relevance ratio: Ratio of matching gold standard diagnoses in the ISABEL list to the total number of diagnostic suggestions offered by ISABEL.

Interrogation time: Time taken to enter clinical data into the system and generate a diagnostic suggestion list (on a 56 KBps modem connection)

Results

Usage statistics

30 GB of data was transferred in the period from July 2001-January 2004 (average/day: 34 MB). There were 10,340,390 successful page requests (average/day: 11,114); over 14,000 users registered to use the site in the specified period. Over a fifth of users accessed the system >5 times since registration. The entire National Health Service (NHS), UK was provided log-in free access via IP address recognition in mid-2002; it then proved difficult to estimate the true number of UK users. This facility was also extended to cover many teaching hospitals in the US (over 10% of the registered users are currently US-based).

Clinical performance of the system

The panel provided gold standard diagnoses for all 100 cases. The median number of such diagnoses per case was 2 (range 1-4). ISABEL provided a maximum number of 15 diagnoses (minimum 10, mode 15). In 73/100 cases, ISABEL displayed *all* gold standard diagnoses (comprehensiveness ratio 1.0). In an additional 15/100 cases, at least *half* of the gold standard was present in the ISABEL suggestions (comprehensiveness ratio 0.50). The mean comprehensiveness ratio across all 100 cases was 0.81.

Since the best raw relevance ratio in this study could only have been 0.27 (all 4 gold standard diagnoses matched in a set of 15 ISABEL diagnoses), a final relevance ratio was calculated (expressed as a proportion of 0.27). In this study, the mean final relevance ratio was 0.45 (95% CI 0.39-0.51).

Over a 56KBps modem connection, ISABEL results took less than 1 sec to display in all cases. Time taken to enter clinical data into the system (interrogation time) was variable depending on the level of detail entered (range: 30 sec – 2 min).

Discussion

This paper describes the development and preliminary analysis of the performance of a diagnostic reminder tool for pediatric medicine. We have shown that, using a novel technique to search an established medical knowledge base, it is possible to deliver relevant diagnostic suggestions in a suitable format for physicians' consideration. The system does not aim to provide probabilistically ranked diagnoses like other similar systems. We feel that organizing diagnostic suggestions in terms of patho-physiological causes is useful for the clinician. Further information on each diagnosis can be sought in the form of text from established medical textbooks. This approach empowers the user, and leaves the final decision making capacity in their hands (treating them as 'learned intermediaries') [15].

Other diagnostic systems have attempted to closely replicate the human processes involved in diagnostic decision making. They were intended to be expert systems, functioning at a level akin to a diagnostic consultant. Human efforts at making medical diagnoses involve, among others, some implicit method of assigning probabilities (a priori, as well as posterior) associated with clinical features, and reconciling patterns learnt or observed from clinical experience. However, this approach has limitations – low base-rate events, which have enormous clinical significance if missed, may be assigned lower probability estimates ('common things are common') during clinical encounters leading to diagnostic 'errors of omission' [16]. Furthermore, such errors may not result always from not knowing, but may be a result of the loss of a checklist function when busy or fatigued during clinical work [17]. It has been demonstrated that using checklists to process many medical tasks leads to improvement in clinical care [18]. In that sense, ISABEL was intended only as a reminder system, to prompt consideration of alternative diagnoses that may have been pushed down by clinicians in their rank order, either because they were uncommon or due to simple omissions.

The development time involved to reach a working prototype of the system was only in the order of months, rather than years. Utilizing established and recognized knowledge bases and applying advanced textual pattern-recognition techniques to the matter contained within them is a novel approach, which ensures minimal manipulation of data by non-experts developing the system. It was apparent during the development of this system that further input by medical experts to fine-tune the raw output of the system was necessary. Previous attempts at developing diagnostic systems have taken many years and involved input from many medical experts. This point is clear from studying the Internist system, that later developed into the QMR system [19]. Updates to the ISABEL system are easy: new text simply replaces the old text in the diagnostic tree. Keeping previously mentioned expert systems up to date was an arduous task that involved searching through the literature for new updates, and consulting with many medical experts.

We used previously researched outcome measures to characterize the clinical performance of the ISABEL system. In comparison to four expert diagnostic systems tested previously by Berner et al in 1994, ISABEL performs well, with a comprehensiveness ratio of 0.81[20]. Testing the relevance ratio was also quite important: it reflects how focused the diagnostic suggestion list was. This is important, because users may reject a system that displays important diagnoses but also provides many other trivial possibilities, detracting from the value of the relevant suggestions. However, since our gold standard consisted of only a few diagnoses that were considered so clinically important that they could not be omitted, as opposed to all possible relevant diagnoses, our relevance ratios were small. For this reason, we also tested the performance of the system with varying numbers of maximum ISABEL diagnostic suggestions. A maximum of 10 diagnostic suggestions retained the comprehensiveness ratio >0.75, and improved the raw relevance ratio to 0.40. We have also shown that these encouraging results were obtained in a clinically negligible amount of time (<2 min), during which the system was interrogated. This is in contrast to other expert systems that may take an average of 22 min to interrogate [21].

Limitations

Limitations of system design include the fact that negative findings cannot be used to influence the diagnostic suggestions produced, and that diagnoses are not ranked in order of probability. These trade-offs were deliberate so that the system remained safe for use. Using negative findings to exclude otherwise important diagnoses may be dangerous considering the uncertainties of clinical medicine. Results from usage statistics suggest that the system is popular, perhaps indicating that users find its advice useful [22], but a more comprehensive survey is needed to further determine usability issues. The clinical performance evaluation described is a preliminary study, and the case mix in the validation sample may not have been fully representative. In addition, results from an isolated examination of the system cannot be extrapolated to suggest the impact of the system on clinicians. It is important that the system's utility is assessed by means of different studies that focus on the impact of the system on diagnostic decision-making.

Conclusions

The ISABEL system promises a novel method of delivering clinically relevant diagnostic suggestions for a variety of clinical scenarios in pediatrics in a negligible amount of time. This model can be extrapolated to develop similar systems for adult medicine.

Acknowledgments

The authors would like to acknowledge the useful comments provided by Dr Jeremy Wyatt and Dr Paul Taylor during the development of the ISABEL system. We also remain thankful to Jason and Charlotte Maude, who were instrumental in setting up the ISABEL Medical Charity to support the development of the ISABEL system.

References

[1] Wyatt JC. *Clinical Knowledge and Practice in the Information Age: a handbook for health professionals.* London: The Royal Society of Medicine Press; 2001.

[2] Wyatt J. Uses and sources of medical knowledge. *Lancet* 1991;338:1368-72.

[3] Brassey J, Elwyn G, Price C, Kinnersley P. Just in time information for clinicians: a questionnaire evaluation of the ATTRACT project. *BMJ.* 2001 Mar 3;322(7285):529-30.

[4] Godlee F, Smith R, Goldmann D. Clinical evidence. *BMJ.* 1999 Jun 12;318(7198):1570-1.

[5] Smith R. What clinical information do doctors need? *BMJ* 1996;313:1062-1068.

[6] Ely JW, Osheroff JA, Ebell MH, Bergus GR, Levy BT, Chambliss ML, Evans ER. Analysis of questions asked by family doctors regarding patient care. *BMJ.* 1999 Aug 7;319(7206):358-61.

[7] Leape LL, Brennan TA, Laird N, Lawthers AG, Localio AR, Barnes BA, Hebert L, Newhouse JP, Weiler PC, Hiatt H. The nature of adverse events in hospitalized patients. Results of the Harvard Medical Practice Study II. *N Engl J Med.* 1991 Feb 7;324(6):377-84.

[8] Barnett GO, Cimino JJ, Hupp JA, Hoffer EP. DXplain: an evolving diagnostic decision-support system. *JAMA.* 1987;258:67-74.

[9] Miller R, Masarie FE, Myers JD. Quick medical reference (QMR) for diagnostic assistance. *MD Comput.* 1986 Sep-Oct;3(5):34-48.

[10] Warner HR Jr. Iliad: moving medical decision-making into new frontiers. *Methods Inf Med.* 1989 Nov;28(4):370-2.

[11] Graber MA, VanScoy D. How well does decision support software perform in the emergency department? *Emerg Med J.* 2003 Sep;20(5):426-8.

[12] Greenough A. Help from ISABEL for paediatric diagnoses. *Lancet.* 2002 Oct 19;360(9341):1259.

[13] Thomas NJ. ISABEL. *Critical Care* 2002; 7(1):99-100.

[14] Ramnarayan P, Britto J. Paediatric clinical decision support systems. *Arch Dis Child.* 2002 Nov;87(5):361-2.

[15] Brahams D, Wyatt J. Decision-aids and the law. *Lancet* 1989;2:632-4.

[16] Bornstein BH, Emler AC. Rationality in medical decision making: a review of the literature on doctors' decision-making biases. *J Eval Clin Pract.* 2001 May;7(2):97-107.

[17] Graber M, Gordon R, Franklin N. Reducing diagnostic errors in medicine: what's the goal? *Acad Med.* 2002 Oct;77(10):981-92.

[18] Balas EA, Weingarten S, Garb CT, Blumenthal D, Boren SA, Brown GD. Improving preventive care by prompting physicians. *Arch Intern Med.* 2000 Feb 14;160(3):301-8.

[19] Miller RA, Pople HE Jr, Myers JD. Internist-1, an experimental computer-based diagnostic consultant for general internal medicine. *N Engl J Med.* 1982 Aug 19;307(8):468-76.

[20] Berner ES, Webster GD, Shugerman AA, Jackson JR, Algina J, Baker AL, Ball EV, Cobbs CG, Dennis VW, Frenkel EP, et al. Performance of four computer-based diagnostic systems. *N Engl J Med.* 1994 Jun 23;330(25):1792-6.

[21] Friedman CP, Elstein AS, Wolf FM, Murphy GC, Franz TM, Heckerling PS, Fine PL, Miller TM, Abraham V. Enhancement of clinicians' diagnostic reasoning by computer-based consultation: a multisite study of 2 systems. *JAMA.* 1999 Nov 17;282(19):1851-6.

[22] Berner ES. Diagnostic decision support systems: how to determine the gold standard? *J Am Med Inform Assoc.* 2003 Nov-Dec;10(6):608-10.

Address for correspondence

Padmanabhan Ramnarayan

Clinical Research Fellow, Imperial College London

ISABEL Medical Charity, Ground Floor, Acrow Building

St Mary's Hospital, Paddington, London, United Kingdom

W2 1NY