Computational Intelligence, Volume 22, Number 3/4, 2006

AN EVALUATION OF THE USEFULNESS OF EXPLANATION IN A CASE-BASED REASONING SYSTEM FOR DECISION SUPPORT IN BRONCHIOLITIS TREATMENT

DÓNAL DOYLE AND PÁDRAIG CUNNINGHAM

Department of Computer Science, Trinity College, Dublin 2, Ireland

PAUL WALSH

Department of Emergency Medicine, Kern Medical Center, Bakersfield, CA 93305, USA

The research presented here explores the hypothesis that the deployment and acceptance of decision support systems in medicine will be enhanced if the basis for the recommendation produced by the system is apparent. We describe a decision support system for advising on patients suffering from bronchiolitis. This system supports its recommendations with *precedent* cases selected to support the recommendation along with justification text that highlights aspects of these cases relevant to the query case. It also presents an estimate of its confidence in the recommendation. The main contribution of this paper is an evaluation of this system in a clinical context. The evaluation shows that this type of explanation does enhance the usefulness of the system for practitioners.

Key words: decision support systems, case-based reasoning, explanations, justification, confidence.

1. INTRODUCTION

It is widely accepted that interpretability and transparency are desirable characteristics of decision support systems (Southwick 1991; Richards 2003). This is particularly the case in medical decision support where the serious consequences that decisions may have require that practitioners have confidence in any decision support system they use. Helping users understand the basis for recommendations is an important step in establishing confidence.

The objective of the research described here is to develop and evaluate case-based explanation systems that are *knowledge-light*. By *knowledge-light* we mean explanation systems where the cases do not contain explicit explanation structures; instead, explanation is achieved by comparison of the query case with retrieved cases. A particular focus of the approach we take is to present cases that support an a fortiori argument, i.e., if a decision were appropriate in one case then it is even more appropriate in another case if the characteristics and symptoms supporting the course of action are even more clear in the second case. For instance, if patient A is similar to our query case and was discharged with a slightly elevated temperature, then the query case who has a lower temperature can also be discharged.

In this article, we present a case-based reasoning (CBR) system that supports its reasoning with explanations of this type. The CBR system is a decision support system for advising on children that are presented at an emergency department (ED) with bronchiolitis. It advises on whether the child can be discharged after treatment or whether the condition is serious enough to warrant admission. In addition to this recommendation the system presents an explanation made up of a previous case that is selected to be the optimum case for explaining the recommendation, justification text highlighting aspects of the query and the selected case and an estimate of the confidence the system has in its recommendation. The main contribution of this article is an evaluation of the elements of the system. This evaluation was carried out in trials of the system at the Kern Medical Center in Bakersfield, CA, in 2005.

We begin with a review in the next section of other CBR systems that have been developed for use in medical domains. The issues associated with determining the disposition of children presenting with bronchiolitis are described in Section 3 and the design of the CBR system is presented in Section 4. An evaluation of this system is described in Section 5 of this paper. The paper finishes with some conclusions in Section 6.

2. RELATED WORK

Since the early work on CASEY (Koton 1988), a CBR system for heart failure diagnosis, CBR has been used in many medical domains. As CBR systems use actual past cases to perform reasoning, these cases can then be used as the basis for explanations. This approach of using precedent cases as an explanation is commonly used by medical professionals. In the Auguste Project, which is concerned with providing decision support for planning the ongoing care of Alzheimer's disease patients, clinicians commented that the system's suggestions were comparable to those a knowledgeable clinician would make (Marling and Whitehouse 2001).

In the CARES system, which is used for predicting the recurrence of colorectal cancer (Ong et al. 1997) doctors are presented with the ten most similar cases to the current query case. This allows the doctor to detect any possible anomalies in the prediction generated by the system. CARE-PARTNER supports the long-term follow-up of cancer patients who have undergone bone marrow transplants (Bichindaritz et al. 1998). This system works in conjunction with rule-based reasoning and information retrieval techniques to perform recommendations.

In CBR systems the most similar case is often the case selected to use as an explanation. In this paper, we will be looking at a technique that we have used to select alternative cases as a basis for generating explanations.

3. PROBLEM DOMAIN

Bronchioles are the noncartilaginous tubules in the pulmonary tree and are less than 1 mm in diameter. Bronchiolitis means inflammation of these bronchioles. In current usage by family practitioners, pediatricians and emergency physicians (EP) bronchiolitis generally refers to a viral illness leading to virally mediated inflammation of the bronchioles in children less than two years of age (Baker and Ruddy 2000). The effect of this inflammation is to narrow the caliber of the bronchioles. This in turn leads to decreased airflow to and from the alveoli (Baker and Ruddy 2000; Orenstein 2000).

Bronchiolitis is common, being the leading cause of hospital admission for this age group in many hospitals (Orenstein 2000). The same viruses that cause bronchiolitis in infants cause little more than a bad cold in older children and adults.

The complications of bronchiolitis arise from airway obstruction, hypoxia, obligate nasal breathing, and the infants' desire to bottle or breast feed. The proportion of lung comprising of smaller airways and the size of these directly relates to the size and age of the child. Younger children are more severely affected. Infants less than two months of age, those with comorbidities, and premature infants are most at risk of respiratory failure and death (Baker and Ruddy 2000).

The clinical picture is of an infant or toddler with runny nose, wheezing, and labored breathing who may be dehydrated. The diagnosis is generally straightforward. Which children should be admitted and which should be discharged is controversial. The decision to admit or discharge is termed disposition. It is one of the most important decisions an EP makes. For many patients the disposition is clear-cut. In marginal cases, being able to draw on the collective outcome experience of the entire ED staff would be desirable. In other cases, it could serve as a safety device by flagging anomalous dispositions in real time.

4. BRONCHIOLITIS SYSTEM

The system was evaluated during the bronchiolitis season in the ED of Kern Medical Center, Bakersfield, CA. This study was approved by Kern Medical Center's institutional review board and informed consent was obtained from the parents of all subjects. When a patient suffering from bronchiolitis was presented to the hospital, the system produced a recommendation to discharge or admit the patient. Section 4.1 will describe the process of producing a recommendation. In addition to the recommendation, an explanation supporting the recommendation was presented to the user. The explanations are made up of an explanation case, justification text, and a measure of the systems confidence in the recommendation. The explanation case is selected to strengthen the argument in favor of the recommendation that has been generated (see Section 4.2). Section 4.3 describes the generation of the justification text which is based on the explanation case while Section 4.4 describes the determination of the systems confidence in its recommendation. Finally in Section 4.5 we describe an overview of using the system and show an example recommendation and explanation from the system.

4.1. Generating Recommendations

CBR is used by the system to generate recommendations. The CBR approach is based on two observations of real-world problem solving. The first is that similar problems tend to have similar solutions, and secondly that the types of problems encountered tend to recur over time. The idea behind CBR is to retrieve and adapt previous cases when solving a problem. Equation (1) shows how to determine the similarity between a target case q and x, a case in the case base, where f is an individual feature in the set of features F, w_f is the weight of the feature f and $\sigma()$ is a measure of the contribution to the similarity from feature f. A classification can then be generated based on the most similar cases.

$$\mathrm{Sim}(q, x) = \sum_{f \in F} w_f \sigma(q_f, x_f). \tag{1}$$

In this system the three most similar cases to a presented target case are considered. Three cases were selected based on a cross validation. If all three of the most similar cases were discharged than the target case is classified as a discharge; otherwise an admission is recommended. This biases the system toward recommending admission.

4.2. Selecting Explanation Cases

In CBR the most similar case is often used as part of an explanation to support a recommendation. However, the most similar case to a target case may not be the most convincing case to support a recommendation. For instance, if a decision is being made on whether to keep a sick twelve-week-old baby in hospital for observation, a similar example with a fourteen-week-old baby that was kept in is more compelling than one with an eleven-week-old baby (based on the notion that younger babies are more likely to be admitted).

We have developed a framework for selecting more convincing cases for use in explanation (Doyle et al. 2004). The framework was implemented using FIONN (Doyle et al. 2005), a Java-based workbench based on CBML (Coyle et al. 2004). Once a classification is performed, the top-ranking neighbors are reranked to explain the recommendation. This ranking is performed using an explanation utility measure shown in (2).

$$\mathrm{Util}(q, x, c) = \sum_{f \in F} w_f \xi(q_f, x_f, c), \tag{2}$$

where $\xi()$ measures the contribution to explanation utility from feature f. The top-ranking case is then selected for use as an explanation case. The utility measure closely resembles the similarity measure (shown in Equation (1)), used for performing the initial nearest-neighbor classification except that the $\xi()$ functions will be asymmetric compared with the corresponding $\sigma()$ functions and will depend on the class label c. This close resemblance maintains the knowledge-light approach of this technique. Examples of utility metrics are shown in Figures 2 and 3 and more information on defining utility metrics can be found in Doyle et al. (2004).[1]

Because the utility measure contains more domain knowledge than the similarity measure, it has been suggested that it should be used to perform the complete classification process rather than just the reranking for explanation. We have investigated this idea but found that it produces a lower classification accuracy than the nearest-neighbor process. Using the utility measure for classification is not completely straightforward as the utility metric is class dependent as shown in equation (2). This can be addressed by using the utility metric to rank the entire case base twice, once for each outcome class. The utility score for the k nearest neighbors for each class is summed and the class with the highest score is returned as the prediction.

To test the effectiveness of this approach to classification, a leave-one-out cross-validation was performed comparing this utility-based classification with the standard similarity-based process. In the three domains in which this approach was tested there was a significant reduction in accuracy using this approach compared to the standard similarity based process. This shows that the requirements for classification accuracy and explanation are different and supports the idea of having an explanation utility framework that is separated from the similarity mechanism used for classification.

4.3. Justification Text

Ashley (1987, 1991), Ashley and Aleven (1997), Mcsherry (2004), and Nugent and Nugent and Cunningham (2004) argue that if a user is to accept a recommendation it is important for them to see pros and cons for that recommendation. We have also developed a technique that uses explanation utility measures to generate justification text that can highlight supporting and nonsupporting features of the explanation case (selected using the technique described in the previous section). When the following inequality holds true we consider that the feature value in the explanation case supports the recommendation:

$$\xi(q_f, x_f, c_p) > \xi(q_f, x_f, c_n)$$

that is, contribution to utility from feature f in the explanation case for disposition c_p is greater than it would be for the opposite disposition c_n. For example, consider a recommendation to discharge a six-month-old patient. The contribution to utility of a four-month-old patient for a discharge disposition is greater than the contribution for a disposition to admit. In a similar manner a feature does not support the recommendation when the opposite inequality holds:

$$\xi(q_f, x_f, c_p) < \xi(q_f, x_f, c_n).$$

Features that are considered to be supportive and nonsupportive of a recommendation are incorporated into justification text as shown in Figure 1.

[1]Full details of the weights and utility metrics used can be found at https://www.cs.tcd.ie/research_groups/mlg/broncexp/.

Features	Patient	Explanation Case
Age (months)	10.2	6.3
Birth	Vaginal	Vaginal
Smoking mother	No	Yes
Hydration before treatment	5% dehydrated	Normal
O_2 saturation before treatment	95	95
Retraction severity before treatment	Mild	Mild
Heart rate after treatment	136	155
Overall increase in work of breathing after treatment	Mild	Mild
Oxygen saturation under 92 after treatment	No (99.0)	No (98.0)
Respiratory rate after treatment	No (30)	No (35)
Temperature over 100.4 after treatment	No (98.0)	Yes (101.0)
Work of breathing after treatment	Improved	Same
Disposition		Discharge

We suggest that this patient should be discharged from hospital.

In support of this recommendation we have the Explanation Case that appears to have been sicker than this patient (due to smoking mother, higher temperature after treatment, and worse work of breathing after treatment) but was still discharged from hospital.

However, it should be noted that the patient's better hydration before treatment in relation to the Explanation Case is a feature that goes against our argument that the explanation case is sicker than the patient.

We have a high confidence in our recommendation.

FIGURE 1. An example recommendation and explanation from the deployed system. In this situation the recommendation is that the patient should be discharged. The explanation is made up of the explanation case (top right), the justification text (bottom half), and finally the level of confidence (last line).

4.4. Confidence

Cheetham and Price have recently emphasized the importance of being able to attach confidence values to recommendations in CBR (Cheetham 2000; Cheetham and Price 2004). This has been a research issue since the earliest days of expert systems research: it is part of the body of research on meta-level knowledge (Lenat et al. 1983; Davis and Buchanan 1985), the view being that it is important for a system to know what it knows.

In Delany et al. (2005), we describe a technique used for determining confidence in a case-based spam filtering system called ECUE. This technique uses an aggregation-based approach to combining individual confidence measures. The bronchiolitis system uses a similar approach to approximating confidence. In this system five different confidence measures are used. The five measures used for confidence are as follows:

Similarity ratio. The similarity ratio calculates the ratio of the sum of similarities between the target case t and its k nearest like neighbors (NLNs) to the sum of similarities between the target case and its k nearest unlike neighbors (NUNs).

Average NUN index. The average nearest unlike neighbor index is a measure of how close the first k NUNs are to the target case t.

Similarity vote. The similarity vote is the sum of the global similarity scores, as calculated using equation (1), for its k nearest neighbors.

Neighbor accuracy. The neighbor accuracy is based on how accurately the neighbours are classified by the system using leave-one-out validation.

Explanation utility ratio. The explanation utility ratio calculates the ratio of the explanation utility between the target case t and its k explanation cases to the utility between the target case and its k NUNs.

If any of the five measures result in a value greater than a predefined threshold the system has a high confidence in its classification; otherwise it has a reasonable confidence. The thresholds are determined using a cross validation process on the training data. The level of confidence is then indicated in the explanatory text. These confidence measures are described in detail in Doyle (2005).

4.5. System Operation

On presentation of a new query case the system generates a recommendation and a supporting explanation. The explanation is made up of an explanation case and justification text. For the most part this text supports the recommendation; however, it also highlights issues that might support a different disposition for the query case to that in the explanation case (see Figure 1). The system operates as follows:

- A new patient suffering from bronchiolitis is entered into the ED's records in the normal way.
- The bronchiolitis system generates a recommendation and supporting explanation (Figure 1).
- The recommendation and explanation are presented to any attending doctors, resident doctors, or nurse practitioners dealing with the patient, who then fill out an evaluation on the recommendation and the case.

Figure 1 shows an example recommendation and explanatory text for a particular patient. The first line of the text shows that the system recommends that the patient should be discharged from hospital. The next paragraph displays the features that it believes supports its recommendation. The third paragraph is included if the system believes that certain features should be considered before a user accepts the recommendation. Finally the system indicates that it has a high level of confidence in the recommendation.

5. EVALUATION

The system was prospectively validated during the months February to April 2005. After completion of patient care, the users were asked to fill out an evaluation for each recommendation and explanation that was produced. The evaluation consisted of three questions:

Question one. Do you agree with the suggested course of action?
Question two. Did you find the explanation case useful?
Question three. Did you find the justification text useful?

Each of these questions had five options to select from; `Definitely Not`, `No`, `Maybe`, `Yes`, and `Absolutely`. Also included in the evaluation was a facility for the user to add any comments they may have. As a result of feedback from these comments, the system was updated halfway through the evaluation period. Therefore, the evaluation was split up into two parts. Section 5.1 describes the initial part of the evaluation and the updates made to the system. Section 5.2 describes the results of the evaluation after the update. Results of a comparison between using a nearest neighbor and a case selected using the explanation utility metrics are presented in Section 5.3. Section 5.4 presents the results from an evaluation on the usefulness of including a counter example as part of an explanation.

5.1. Results: Part 1

This part of the evaluation ran until mid March 2005. The final results closely match the results found in the second part. For this reason we will not elaborate much on the results in this phase. However, the main finding in this case was that the age of the explanation case often differed greatly from that of the patient. In the comments the users often suggested that a less marginal case with respect to the patient would be more useful.

These comments suggested that we should modify the shape of the utility graph for the age feature to reduce the tendency to invoke extreme example cases. The original utility measure for age when the classification is `discharge` is shown in Figure 2. In this scenario consider a query case, q, with `Age` equal to fifteen months and a retrieved case, x, with `Age` equal to ten months. The difference between these two values $(q - x)$ is $+5$. By looking up the graph it can be seen that a difference of $+5$ returns an explanation utility of 1.

A utility of 1 is high for an age difference of $+5$ months if we wish to retrieve cases that are closer in age to the patient. Figure 3 shows the updated utility metric. Here we can see that the difference of $+5$ has now dropped, while smaller differences remain high. More information on these difference measures can be found in Doyle et al. (2004).

5.2. Results: Part 2

The second part of the evaluation used the updated utility measures as described in the previous section and ran from mid-March until the end of April. In total, out of sixty-five recommendations, the system had a high confidence twelve times. For each of these confident recommendations the user also either agreed or strongly agreed with the recommendation.

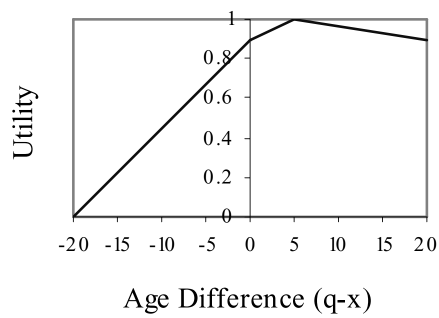

FIGURE 2. Original age explanation utility measure for a discharge disposition.

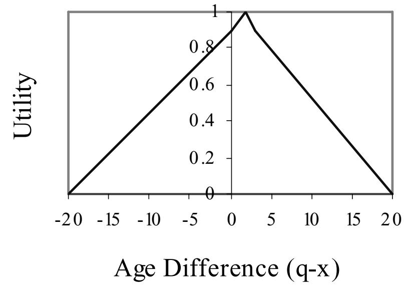

FIGURE 3. Updated age explanation utility measure for a discharge disposition.

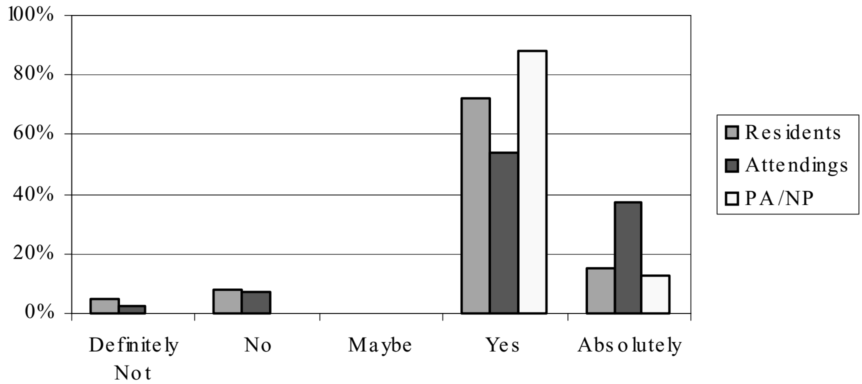

FIGURE 4. Q1: Do you agree with the suggested course of action?

However, when the recommendation of each of these cases was compared to the final disposition,[2] the recommendation was correct ten out of the twelve times.

For the 65 recommendations in this part, there were 106 evaluations; 39 from residents, 59 from attendings, and 8 from physician assistants/nurse practitioners (PAs/NPs).[3] In question one we looked at the user's overall confidence in the recommendation generated by the system. Figure 4 shows that over 85 percent of each group had either some confidence or total confidence in the recommendation.

In question two, we were checking to see if the users found the explanation case to be a suitable case for explaining the recommendation. The results for this question are shown in Figure 5. Here we can see that 54 percent of residents, 56 percent of attendings, and 100 percent of PAs/NPs found the explanation case to be useful. It is the evaluations for this question from the attendings that had the most significant change when compared to the first part of the evaluation. In the first part, 44 percent of attendings answered `yes` to this question, while only 2 percent answered `absolutely`, while in the second part this increased to 49 percent and 7 percent, respectively. Although it could be argued that this increase could be partially attributed to an increase in accuracy of the system from 73 percent to 80 percent, it appears that the increase is mainly due to patients of a more similar age being used as

[2]If a patient is discharged, a follow-up call is made three days later to make sure the patient was not subsequently admitted to hospital with bronchiolitis. If the patient was subsequently admitted, their final disposition is an admission.

[3]Attendings have the highest level of expertise. PAs/NPs have less training but more experience than residents.

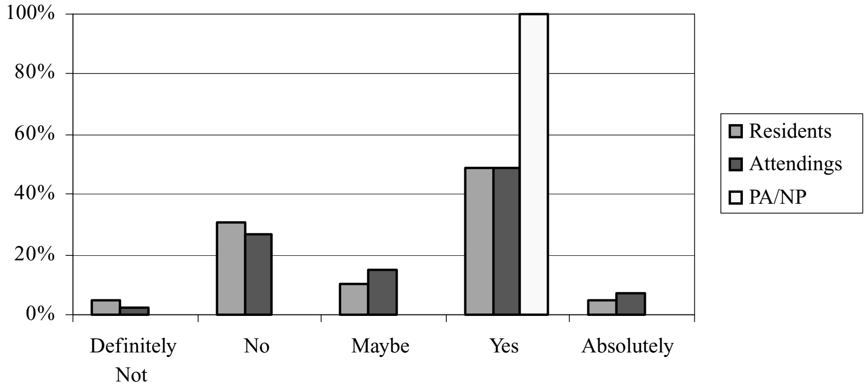

FIGURE 5. Q2: Did you find the explanation case useful?

explanation cases. This is shown by a substantial reduction in comments complaining about an excessive age gap. As more cases are added to the case base, this problem should be completely eliminated.

Finally in question three we are interested in the quality of the generated justification text. Figure 6 shows the results for this part of the evaluation. As with the explanation case the majority was happy with the justification text.

5.3. Explanation Case and Nearest-Neighbor Comparison

The next stage of the evaluation was to perform a direct comparison of the explanation case, selected using the technique described in Section 4.2, against the use of the nearest-neighbor to support a recommendation. To do this, an off-line evaluation was performed using the bronchiolitis data. Subjects from the Kern Medical Center were presented with ten unique problem cases and associated nearest-neighbor and explanation case—labeled randomly as explanation 1 and explanation 2. Out of the ten problem cases selected five had recommendations to discharge and five had recommendations to admit and these were presented in a random order. The subjects were asked to select the case that they felt was

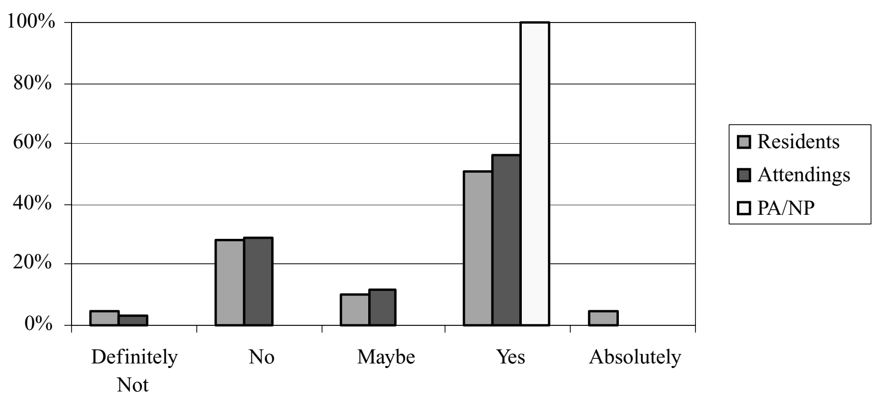

FIGURE 6. Q3: Did you find the supporting explanatory text useful?

TABLE 1. Summary Results of Explanation Case vs. Nearest Neighbor in the Bronchiolitis and Breathalyzer Domain

Domain	Explanation Case	Nearest Neighbor	Neither	Either
Bronchiolitis	32	72	7	29
Breathalyzer	96	18	5	11

most supportive of the recommendation. They also had the option to state that they found the cases to be equally supportive or to state that neither case was supportive.

The evaluation was performed by fourteen subjects. Table 1 shows a summary of the results from this evaluation. In this situation we can see that the nearest neighbor was preferred as an explanation seventy-two times compared with thirty-two preferences for the case selected using the explanation utility framework. This is quite surprising as the opposite was found to be true in two other domains. The use of the explanation framework in the Breathalyzer domain is described in Doyle et al. (2004) and a further evaluation in a diabetes e-clinic application is described in Doyle et al. (2005). The cases selected by the explanation framework are strongly favored in both these applications. The results of the evaluation in the Breathalyzer application is presented in Table 1—it can be seen that the explanation case is five times more popular than the nearest neighbor in this situation. The Breathalyzer application contains five features and is used for predicting if a subject is over or under the legal blood alcohol limit for driving in Ireland (Cunningham et al. 2003). The layout of this evaluation was the same as the evaluation used in the bronchiolitis application. Again ten unique problem cases were selected; five with classifications of under the legal limit and five with classifications of over the legal limit. In total thirteen subjects from the Computer Science Department of Trinity College Dublin performed the evaluation. The time spent per question was recorded to help assess the attention paid by subjects to the evaluation. The average time completing the evaluation was over ten minutes. Any users spending less than five minutes performing the evaluation were removed from the evaluation.

We believe that the difference in performance is accounted for by the high number of *directional features* in the bronchiolitis problem (10) compared with three and four in the e-clinic and Breathalyzer applications. By *directional features* we mean features of a case that can be used as part of an a fortiori argument. In the simpler e-clinic and Breathalyzer situations the influence of these directional features is apparent and the a fortiori advantages of the explanation case are apparent. This is less apparent in the bronchiolitis application because of the larger number of features that influence the outcome.

Thus, of the two benefits that the explanation utility framework offers (name by case selection and feature highlighting), the ability to select useful explanation cases is not manifest in the bronchiolitis application. The ability to select features for highlighting in the justification text is still there however. In future work we plan to modify the explanation framework so that the full case description is not used in complex problems such as the bronchiolitis application. Instead, the explanation would be based on a reduced set of features that would be selected using the explanation utility criteria.

5.4. Counterexample Evaluation

The explanation utility measures attempt to retrieve cases that lie between a problem case and the decision boundary. A possible improvement to just displaying the explanation case is to also display a counter example. The HYPO system (Ashley 1989) contests legal arguments by also citing a past case as a counterexample. The motivation for showing the counterexample is to give the user a sense of the "robustness" of the classification. If the nearest counterexample is quite different to the query case, then the user can have some confidence that the recommendation is correct. If the counterexample is close to the query, then classification may be more marginal. The counterexample could play an important role when classification is incorrect as it could show that the situation is marginal. However, one problem with displaying the counterexample is that it adds to the amount of information a user has to process when presented with a recommendation and explanation. This can cause information overload for the user and end up confusing the user.

To test the usefulness of displaying a counterexample, we set up an evaluation using the bronchiolitis domain. The evaluation was made up of ten unique problem cases; five with correct recommendations and five with incorrect recommendations presented in random order. For each of the problem cases, subjects were presented with the problem case, recommendation, explanation case, and counterexample. The counterexample was selected as the most similar case with a different classification to the recommendation. The evaluation also consisted of two questions:

Question one. Do you consider the recommendation to be correct?
Question two. Do you think the counterexample was useful?

Each of these questions was scored on a five-point scale using five options to select from: `Definitely Not`, `No`, `Maybe`, `Yes`, and `Absolutely`. Also included in the evaluation was a facility for the user to add any comments they may have and the ability to backtrack to change their score. In the evaluation of the results these scores were interpreted as numeric values from 1–5. The subjects were staff from the ED of Kern Medical Center. The time spent per question was recorded to help assess the attention paid by subjects to the evaluation.

In total twelve subjects performed the evaluation. Table 2 shows the average ratings on a scale of 1–5 for both questions amongst the twelve subjects for the correct and incorrect recommendations. The results from question one show that users had a good understanding of the accuracy of the recommendations. The average rating for correct recommendations is 4.6 out of a possible 5. On the other hand, incorrect recommendations had an average rating of 2.9. The results from question two present a more interesting picture. The average rating for the incorrect recommendations is higher than for correct recommendations. This shows that users found the inclusion of the counterexample more useful when they believed that the case had an incorrect recommendations than when it had a correct recommendation. In spite of the subjects finding the counterexample useful for detecting incorrect recommendations, overall they did not find the counterexample very useful. This is shown by the poor overall ratings for question two where the average value of usefulness of the counterexample for incorrect recommendations of 3.3 equating to a slightly positive maybe and the average rating for correct recommendations of 2.9 equating to a slightly negative maybe.

TABLE 2. Results from the Evaluation of the Usefulness of Using a Counterexample in Explanations

	Q1	Q2
Average rating correct recommendations	4.6	2.8
Average rating incorrect recommendations	2.9	3.3

6. CONCLUSIONS

The research reported here is motivated by the belief that the potential to produce explanations based on precedent is one of the great advantages of CBR. We have presented a knowledge-light framework that captures the idea of explanation utility that selects cases and highlights aspects of those cases for use in explanation. The main contribution of this paper is a clinical evaluation of this explanation framework in a case-based decision support system for advising on children suffering from bronchiolitis. The key question to be decided is whether the child should be admitted to hospital or discharged after treatment. The framework is based on the a fortiori principle: an example of the application of this principle in the bronchiolitis context is that a child might be discharged after treatment if a child with similar but more severe symptoms was also discharged.

The explanation framework has two advantages: it selects cases based on this a fortiori principle that provide strong justifications for a course of action and it highlights aspects of those cases that should be considered. We have demonstrated in earlier work that the explanation utility framework selects explanation cases that are better than the nearest neighbor in the breathalyzer and e-clinic domains (Doyle et al. 2004; Doyle 2005). The evaluation presented here shows that this is not the case in the bronchiolitis domain. We believe this is because of the increased complexity of the bronchiolitis case representation compared with the breathalyzer and e-clinic cases. The bronchiolitis cases have ten directional features whereas the e-clinic and breathalyzer cases have three and four directional features, respectively. It seems clear that the advantage of the explanation utility criteria are not manifest in this more complex domain. In future work we plan to explore the idea of presenting the explanation cases as a reduced set of key features, thus making the directional effects more apparent.

The other advantage of the explanation utility framework is the ability to select important features for highlighting in the explanation text that forms part of the explanation. The clinical evaluation shows that this explanation text is found to be useful.

The final aspect of the evaluation presented here is a consideration of the usefulness of counterexamples in explanation. Our conclusion is that this does not add value because any benefits the counterexample might have in situations where the system is incorrect is outweighed by the doubt raised in the majority of situations where the system is correct.

We feel the evaluation presented here shows the considerable potential of knowledge-light explanation in CBR systems for medical decision support. We feel there is still work to be done in deciding how best to present the details of the explanation cases in complex domains.

REFERENCES

ASHLEY, K. 1991. Reasoning with cases and hypotheticals in hypo. International Journal of Man-Machine Studies, **34**: 753 –796.

ASHLEY, K., and V. ALEVEN. 1997. Reasoning symbolically about partially matched cases. *In* Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence (IJCAI-97). Morgan Kaufmann: San Francisco, Nagoya, Japan, pp. 335–341.

ASHLEY, K. D. 1987. Modelling legal argument: Reasoning with cases and hypotheticals. PhD thesis, Department of Computer and Information Science, University of Massachusetts.

ASHLEY, K. D. 1989. Defining salience in case-based arguments. *In* Eleventh International Joint Conference on Artificial Intelligence (IJCAI-89). *Edited by* N. Sridharan. Morgan Kaufmann, San Mateo, CA, pp. 537–542.

BAKER, M., and R. RUDDY. 2000. Pulmonary emergencies. *In* Textbook of Pediatric Emergency Medicine, (4 ed). *Edited by* G. Fleischer and S. Ludwig. Lippincott Williams & Wilkins, Philadelphia, pp. 1067–1087.

Bichindaritz, I., E. Kansu, and K. M. Sullivan. 1998. Case-based reasoning in care-partner: Gathering evidence for evidence-based medical practice EWCBR '98. Proceedings of the 4th European Workshop on Advances in Case-Based Reasoning, Springer-Verlag, London, pp. 334–345.

Cheetham, W. 2000. Case-based reasoning with confidence. *In* Advances in Case-Based Reasoning, 5th European Workshop, EWCBR 2000, Trento, Italy, September 6-9, 2000, Proceedings, Vol. 1898 of Lecture Notes in Computer Science. *Edited by* E. Blanzieri and L. Portinale. Springer, London, pp. 15–25.

Cheetham, W., and J. Price. 2004. Measures of solution accuracy in case-based reasoning systems. *In* Advances in Case-Based Reasoning, 7th European Conference, ECCBR 2004, Madrid, Spain, August 30–September 2, 2004, Proceedings, Vol. 3155 of Lecture Notes in Computer Science. *Edited by* P. Funk and P. A. González-Calero. Springer, London, pp.106–118.

Coyle, L., D. Doyle, and P. Cunningham. 2004. Representing similarity for CBR in XML. *In* Advances in Case-Based Reasoning, 7th European Conference on Case-Based Reasoning, ECCBR 2004, Vol. 3155. *Edited by* P. A. G. Calero and P. Funk. Springer, Madrid, pp. 119–127.

Cunningham, P., D. Doyle, and J. Loughrey. 2003. An evaluation of the usefulness of case-based explanation. *In* Case-Based Reasoning Research and Development, 5th International Conference on Case-Based Reasoning, Vol. 2689 of Lecture Notes in Computer Science, Springer, ICCBR 2003. *Edited by* K. D. Ashley and D. G. Bridge. Trondheim, Norway, pp. 122–130.

Davis, R., and B. Buchanan. 1985. Meta level knowledge. *In* Rule-Based Expert Systems. *Edited by* F. Hayes-Roth, D. A. Waterman, and D. B. Lenat. Addison-Wessley, London, pp. 507–530.

Delany, S. J., P. Cunningham, D. Doyle, and A. Zamolotskikh. 2005. Generating estimates of classification confidence for a case-based spam filter, to appear 5th International Conference on Case-Based Reasoning, Chicago, IL, pp. 157–168.

Doyle, D. 2005. A knowledge-light mechanism for explanation in case-based reasoning. PhD thesis, Trinity College Dublin, Department of Computer Science. Available at: https://www.cs.tcd.ie/publications/tech-reports/reports.05/TCD-CS-2005-71.pdf.

Doyle, D., P. Cunningham, D. Bridge, and Y. Rahman. 2004. Explanation oriented retrieval. *In* Advances in Case-Based Reasoning, Proceedings of the Seventh European Conference on Case-Based Reasoning. *Edited by* P. Funk and P. Calero. Springer, pp. 157–168.

Doyle, D., J. Loughrey, C. Nugent, L. Coyle, and P. Cunningham. 2005. Fionn: A framework for developing CBR systems. Expert Update, **8**(1):11–14.

Koton, P. 1988. Reasoning about evidence in causal explanations. *In* Proceedings of AAAI-88, St. Paul, MN, pp. 256–261.

Lenat, D., R. Davis, J. Doyle, M. Genesereth, I. Goldstein, and H. Schrobe. 1983. Reasoning about reasoning. *In* Building Expert Systems. *Edited by* F. Hayes-Roth, D. A. Waterman, and D. B. Lenat. Addison-Wesley, London, pp. 219–239.

Marling, C., and P. Whitehouse. 2001. Case-based reasoning in the care of Alzheimer's disease patients, ICCBR '01: Proceedings of the 4th International Conference on Case-Based Reasoning, Springer-Verlag, London, pp. 702–715.

McSherry, D. 2004. Explaining the pros and cons of conclusions in CBR. *In* Advances in Case-Based Reasoning, 7th European Conference, ECCBR 2004 Madrid, Spain, August 30–September 2, 2004, Vol. 3155 of LNAI. *Edited by* P. A. G. Calero and P. Funk. Springer, pp. 317–330.

Nugent, C., and P. Cunningham. 2004. A case-based explanation system for 'black-box' systems. *In* ECCBR 2004 Workshop Proceedings. *Edited by* P. Cunningham and D. McSherry, pp. 155–164.

Ong, L., B. Sheperd, C. L. Tang, F. Seow-Choen, Y. Ho, L. Tong, Y. S. HO, and K. Tan. 1997. The colorectal cancer recurrence support (cares) system. Artificial Intelligence in Medicine **11**(3):175–188.

Orenstein, D. 2000. Bronchiolitis. *In* Nelsons Textbook of Paediatrics. *Edited by* R. Berhman, R. Kliegman, and R. Jenson. Saunders, Philadelphia, pp. 1285–1287.

Richards, D. 2003. Knowledge-based system explanation: The ripple-down rules alternative. Knowledge and Information Systems, **5**(1):2–25.

Southwick, R. R. 1991. Explaining reasoning: An overview of explanation in knowledge-based systems. The Knowledge of Engineering Review **6**(1):1–19.