Comparison of the Evaluations of a Case-Based Reasoning Decision Support Tool by Specialist Expert Reviewers with Those of End Users

Paul Walsh, MB, BCh, BAO* * Kern Medical Center, Department of Emergency Medicine
Donal Doyle, PhD† † University College Dublin
Kemedy K. McQuillen, MD‡ ‡ Central Maine Medical Center
Joshua Bigler, MD# # University of Nevada
Caleb Thompson, MD*
Ed Lin, MD*
Padraig Cunningham, PhD†

Supervising Section Editor: Christian D. McClung, MD
Submission history: Submitted May 8, 2007; Revision Received December 6, 2007;Accepted January 25, 2008.
Reprints available through open access at www.westjem.org

Background: Decision-support tools (DST) are typically developed by computer engineers for use by clinicians. Prototype testing DSTs may be performed relatively easily by one or two clinical experts. The costly alternative is to test each prototype on a larger number of diverse clinicians, based on the untested assumption that these evaluations would more accurately reflect those of actual end users.

Hypothesis: We hypothesized substantial or better agreement (as defined by a κ statistic greater than 0.6) between the evaluations of a case based reasoning (CBR) DST predicting ED admission for bronchiolitis performed by the clinically diverse end users, to those of two clinical experts who evaluated the same DST output.

Methods: Three outputs from a previously described DST were evaluated by the emergency physicians (EP) who originally saw the patients and by two pediatric EPs with an interest in bronchiolitis. The DST outputs were as follows: predicted disposition, an example of another previously seen patient to explain the prediction, and explanatory dialog. Each was rated using the scale Definitely Not, No, Maybe, Yes, and Absolutely. This was converted to a Likert scale for analysis. Agreement was measured using the κ statistic.

Results: Agreement with the DST predicted disposition was moderate between end users and the expert reviewers, but was only fair or poor for value of the explanatory case and dialog.

Conclusion: Agreement between expert evaluators and end users on the value of a CBR DST predicted dispositions was moderate. For the more subjective explicative components, agreement was fair, poor, or worse.
[*West*JEM. 2008;9:74-80.]

INTRODUCTION

Decision-support tools (DST) are typically developed by computer engineers who rely heavily on feedback from clinicians as they build and test the DST prototypes. Often developers will collaborate with one or two clinicians with a particular expertise in the field for which the DST is being targeted. An alternative approach is to test each prototype on a larger number of diverse clinicians, anticipating that these evaluations of the evolving DST will more accurately reflect those of actual end users. This latter approach is logistically far more difficult than the former, adding time and expense to the development of DSTs. Furthermore, the underlying assumption that testing a DST on a larger number of clinicians is better is an untested one.

To test this assumption we compared the evaluations of a DST performed by two sub-specialists with a particular interest in the field targeted by the DST with those of 12 other clinicians. We did this using a case-based reasoning (CBR) tool designed to predict the disposition of children with bronchiolitis. This DST provides three distinct outputs. It predicts disposition. It provides an example of a previously treated patient and that patient's outcome from a database of previously treated patients as evidence supporting its prediction. It provides an explanatory dialog to 'explain' its decision.

We hypothesized substantial or better agreement (as defined by a κ statistic greater than 0.6) between the evaluations of the DST performed by the clinically diverse end users, to those of two sub-specialist reviewers who evaluated the same DST output.

METHODS

The study was approved by our Institutional Review Board. A CBR tool for use in infants with bronchiolitis was developed and prospectively tested in an academic emergency department. This has been described in detail elsewhere.[1] Briefly, the DST compares the patient presented to it with previous patients in its database. It uses nine clinical features, including response to treatment to match the patient as closely as possible to a previous patient for whom the clinical outcome is known. These clinical features are shown in the appendix, which gives a sample DST output. Based on a previously treated patient in the database whose outcome is known, the DST predicts the current patient's disposition. It then presents the case from its database that most supports the prediction and generates a dialog comparing and contrasting the previously seen patient and the current patient.

Following enrollment, a detailed history and physical exam was performed on each child and the results recorded on a specifically mandated data-collection sheet. This information was entered in a customized Filemaker-pro database.[2] The DST extracted the data points it required directly from this database. To prevent the DST from influencing a clinician's disposition decisions, we delayed presenting the DST printed output until the clinician's next shift. Each physician was asked to rate the usefulness of each of the three components of DST output (disposition, case to justify the disposition, and explanatory dialog) using the scale Definitely Not, No, Maybe, Yes, and Absolutely. This was converted to a Likert Scale from 1 to 5 respectively for analysis. An example of this output is shown in Appendix 1. We also analyzed the data compressing the ordinal five-point Likert scale to a three-point scale, as two people could mean nearly exactly the same thing by 'No' and 'Definitely not.'

Two pediatric EPs, both of whom have previously published research on bronchiolitis, acted as the expert reviewers. These experts reviewed the data collection sheet DST output and the CBR-DST output in the same manner as the original end users.

One of these reviewers also performed a blinded review of the cases without the DST output to provide some measure of disagreement that could be attributed solely to the use of chart review rather than due to disagreement with the DST output. This was performed four months before review of the DST output to minimize recall bias.

Severity of illness of the patients was calculated using the NCH bronchiolitis severity model.[3] A predominance of mildly or severely ill patients would render a DST less useful and potentially could affect physicians' evaluations of it.

Inter rater agreement was calculated using a weighted kappa (κ) statistic. The κ statistic was interpreted as recommended by Landis and Koch.[4] Confidence intervals for the weighted κ statistic were calculated using a bootstrap technique.[5] Mean scores, their distribution and interquartile ranges (IQR) for the end users and the expert evaluators, were calculated. Overall distributions of scores were compared using the non parametric sign rank test. Statistical analysis was performed using Stata 9.2 software.

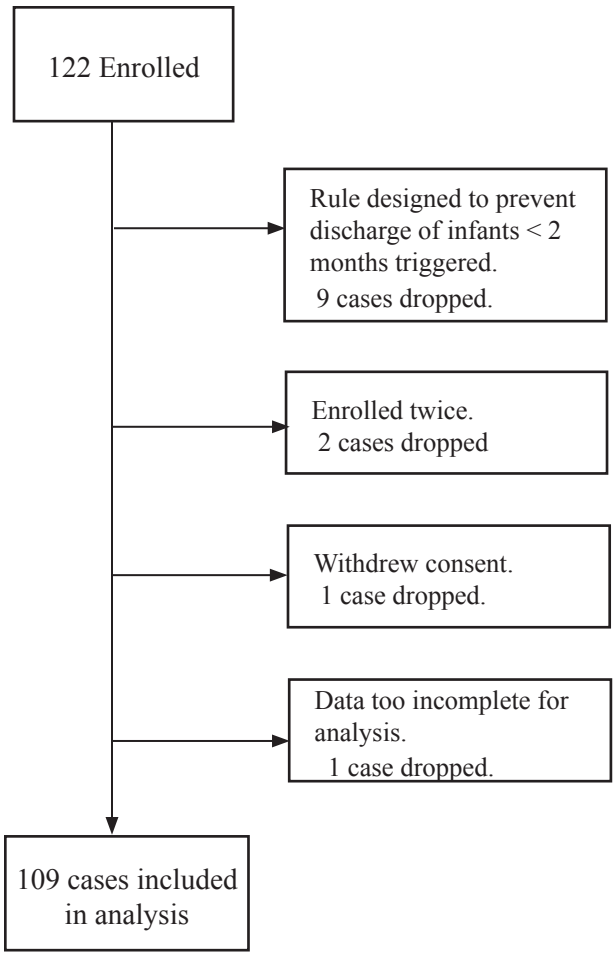

Figure 1. Case flow through the study.

Table 1. Agreement between evaluators on the predicted disposition. The values in parentheses are the results obtained when the five categories are collapsed to three.

CBR DST predicted disposition: Do you agree with the suggested course of action?					
Evaluator 5 point scale (3 point scale)	**Observed Agreement**	**Agreement expected by chance alone**	**κ**	**95% C.I.**	**Interpretation**
End users & Expert 1	93.5% (89.9%)	87.2% (79.6%)	0.49 (0.51)	0.25 - 0.69 (0.25 - 0.71)	Moderate (Moderate)
End users & Expert 2	93.6% (91.6%)	86.4% (79.9%)	0.53 (0.58)	0.33 - 0.68 (0.36 - 0.76)	Moderate (Moderate)
Expert 1 & Expert 2	94.5% (91.6%)	87.3% (80.9%)	0.56 (0.56)	0.38 - 0.70 (0.33 - 0.74)	Moderate (Moderate)

Table 2. Agreement between evaluators on the value of the explanatory case. The values in parentheses are the results obtained when the five categories are collapsed to three.

CBR DST explanatory example: Did you find the explanation case useful?					
Evaluator 5 point scale (3 point scale)	**Observed Agreement**	**Agreement expected by chance alone**	**κ**	**95% C.I.**	**Interpretation**
End users & Expert 1	89.2% (70.8%)	83.0% (58.4%)	0.36 (0.30)	0.19 - 0.53 (0.14 - 0.46)	Fair (Fair)
End users & Expert 2	83.96% (46.0%)	84.3% (43.2%)	-0.02 (0.05)	-0.10- 0.04 (0.04 - 0.14)	None* (Poor)
Expert 1 & Expert 2	87.03% (59.2%)	87.3% (52.6%)	-0.01 (0.14)	-0.08 - 0.07 (0.03 - 0.26)	None* (Poor)

Table 3. Agreement between evaluators on the value of the explanatory dialog. The values in parentheses are the results obtained when the five categories are collapsed to three.

CBR DST explanatory dialog: Did you find the supporting dialog useful?					
Evaluator 5 point scale (3 point scale)	**Observed Agreement**	**Agreement expected by chance alone**	**κ**	**95% C.I.**	**Interpretation**
End users & Expert 1	87.2% (66.0%)	83.6% (56.7%)	0.21 (0.21)	0.03 - 0.40 (0.05 - 0.38)	Fair (Fair)
End users & Expert 2	84.1% (62.0%)	83.4% (58.7%)	0.04 (0.08)	(0.13 - 0.22) (-0.10 - 0.26)	Poor (Poor)
Expert 1 & Expert 2	78.3% (60.3%)	79.2% (56.9%)	-0.04 (0.08)	-0.20 - 0.12 (-0.09 - 0.25)	None (Poor)

RESULTS

One hundred and twenty-two patients were enrolled. Patient flow and exclusions are shown in Figure 1. Following exclusions, 109 remained in the analysis. Expert reviewer evaluations were available for all of these. Attending physicians performed end-user evaluations on 97 of the CBR predictions of disposition and 96 of the CBR explanatory dialogs. Midlevel providers performed the evaluations on 12 cases, three of which had no attending evaluations.

The mean number of years following residency training for the faculty was nine (range one to 28). All were board prepared or certified and all but one residency trained in emergency medicine.

Severity of illness and age characteristics are shown in

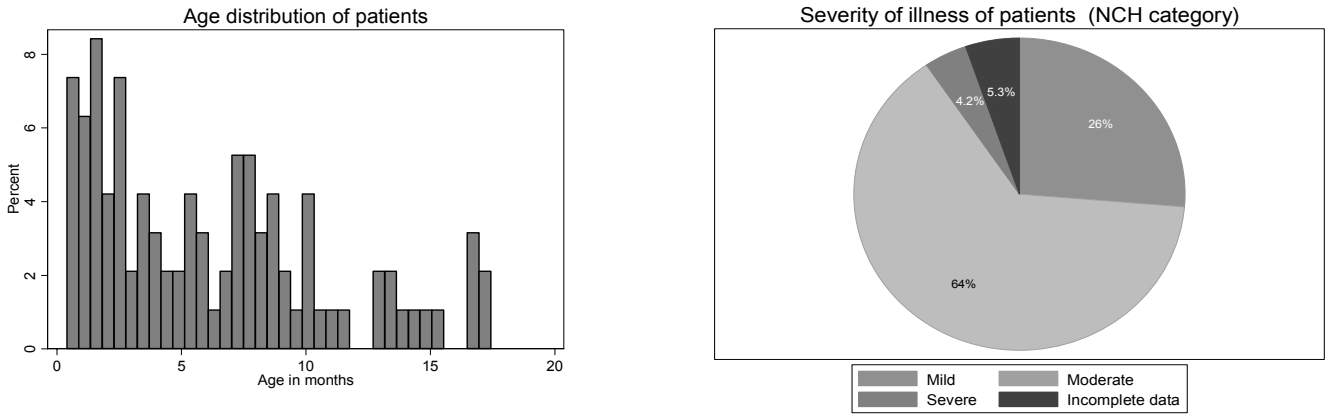

Figure 2. Age and severity of illness of patients.

Figure 2 and showed a broad range of cases. The expert reviewer who performed the chart review agreed with the disposition of the end user 93/109 (85.3%) of the time (expected by chance alone 50%) $\kappa = 0.66$ (95% CI 0.53 to 0.80) demonstrating moderate agreement.

The raw scores and their distribution for the evaluations are shown in Figure 3. Agreement between the end users and expert reviewers and the reviewers with each other are shown in Tables 1 to 3.

DISCUSSION

We found moderate agreement between our expert reviewers and actual end users for disposition, but this decreased progressively as the inherent subjectivity of the DST output being evaluated increased. The expert reviewers did not agree any more with each other than they did with the end users when the case was more ambiguous. For DST developers this is disheartening as it suggests that when developing these tools prototype testing requires feedback from a group representative of actual end users rather than one or two interested clinical experts. This former approach to DST development is logistically much more difficult and costly to perform than the latter. The silver lining for developers was that the end users consistently scored the DST more highly than did the expert reviewers.

LIMITATIONS

The management of bronchiolitis is inherently controversial,[6-8] and some disagreement between clinicians on disposition is to be expected leading to an immeasurable random bias towards poorer agreement in disposition and presumably DST output. On the other hand, it is precisely for such less than clear-cut conditions that CBR may offer some benefit. The use of chart review by the expert reviewers introduces potential bias to decreased agreement. Eliminating this systematic bias would require that the patients were independently seen by both the treating clinician and the expert reviewer at the same time. Such a methodology is unlikely to be feasible in emergency medicine. We addressed this by having one expert perform an initial blinded review of disposition, at least providing some measure of the effect of this. Agreement for this was 85% (compared with 50% expected by chance alone), suggesting that this effect was relatively modest. It implies a methodologically introduced potential upper limit of substantial agreement (κ=0.6 to 0.8) for what might be obtained between the end users and expert reviewer by virtue of the use of chart review by the experts. This is important; the observed agreement was moderate (κ=0.49) for the DST-predicted disposition, suggesting that for this outcome at least the agreement may be better than it appears after initial review.

The role of chance in leading to artificially poor agreement between DST users is minimized by using larger numbers of patients and clinicians. While the number of clinicians involved in the study is relatively small, it strikes a balance between having too few evaluations carried out by many evaluators and the feasibility of obtaining a reasonable sample size of patients. This was particularly the case for our study, which required written informed parental consent for every patient.

The generalizability of our work is limited because we considered a single DST at a single site. We have previously noted a higher discharge rate (with more discharge failures) at our site compared with a second ED. We addressed this in part by using an expert reviewer from another center. While preferable to using a single reviewer from the study site, this likely further decreased agreement. A potential confounder arises from using experts in a field to evaluate a decision support tool. By virtue of their expertise they may find

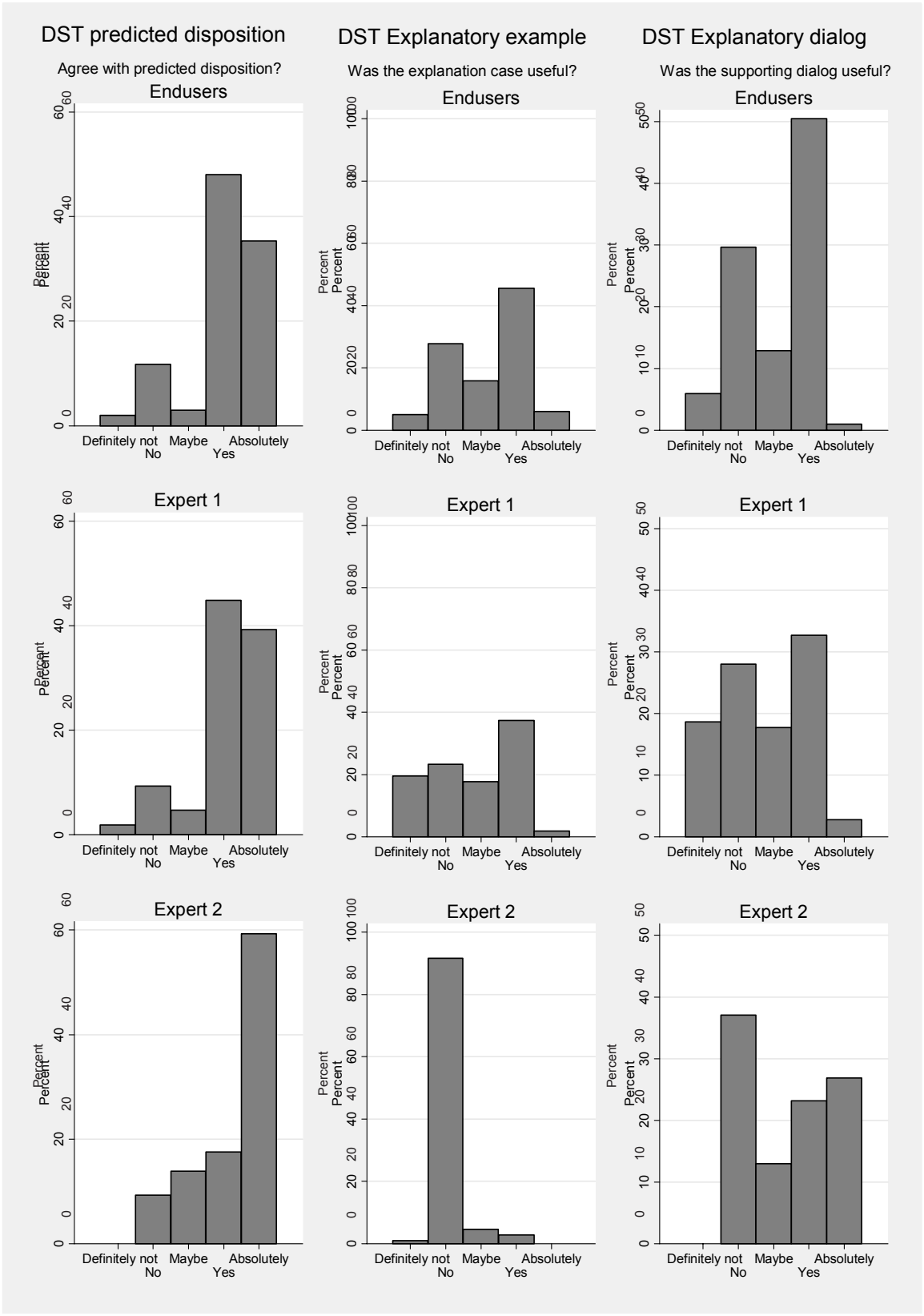

Figure 3. Evaluation (raw scores) of the DST by the end users and expert reviewers.

any DST less useful than their more generalist colleagues, and there is some evidence in this study pointing to this. However, there were few cases according to these reviewers where their opinion of the DST would have been changed regardless of whether they rated its output for their own use or what they perceived as appropriate for a more general audience.

Other potential confounding factors can be missed. Local limitations on bed availability, proximity of the patient's residence to the hospital, and the reliability of the parents affect dispositions in ways unmeasured by the decision support tool we tested. While the arrival of a clearly intoxicated parent will likely be documented, subtler yet important considerations may not be recorded on the patient's chart. For instance, dirty maternal fingernails have been associated with increased infant dehydration[9] but are not often noted on a child's chart. Moreover, estimating the magnitude of the effect of such variables on disposition decisions is difficult. The DST will not reflect these considerations, so such cases will tend to decrease agreement between end users and a subsequent reviewer on the correctness of the DST output. All these considerations tend to make our estimate more conservative. This lends support to using expert reviewers for objective criteria; however, even if our estimate of agreement on subjective DST output is overly conservative, this agreement was so weak that it seems difficult to justify their use.

Answering our question in the general will require replicating experiments like ours with a variety of DST types in various clinical settings for a variety of clinical conditions. In the meantime DST developers must at least consider the implications of this work when prototype testing DSTs.

CONCLUSION

Agreement between expert evaluators and end users with predicted disposition for children with bronchiolitis by a CBR-based DST predicted was moderate. For the more subjective explicative components of the DST output, agreement was fair, poor, or worse. The general clinical end users ranked the DST more highly than the specialist clinical reviewers.

Address for correspondence: Paul Walsh, M.D., Kern Medical Center, 1830 Flower Street, Bakersfield, CA 93305, E-mail: yousentwhohome@yahoo.com

REFERENCES

1. Doyle D, Cunningham P, Walsh P. An evaluation of the use of explanation in a case based reasoning system for decision support in bronchiolitis treatment. *Computational Intelligence*. 2006; 22:269-281
2. Filemaker Pro. Microsoft XP. Santa Clara, CA: Filemaker Inc., 2006
3. Walsh P, Rothenberg SJ, O'Doherty S et al. A validated clinical model to predict the need for admission and length of stay in children with acute bronchiolitis. *Eur J Emerg Med*. 2004; 11:265-272.
4. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977; 33:159-174.
5. Reichenheim ME. Confidence intervals for the kappa statistic. *The Stata Journal*. 2005; 4:421-428. Software component available for download from http://stat-journal.com/software/sj4-4. Accessed 10-12-2007.
6. Baldwin RL, Green JW, Shaw JL et al. Physician risk attitudes and hospitalization of infants with bronchiolitis. *Acad Emerg Med*. 2005; 12:142-146.
7. Bordley WC, Viswanathan M, King VJ et al. Diagnosis and testing in bronchiolitis: a systematic review. *Arch Pediatr Adolesc Med*. 2004; 158:119-126.
8. Christakis DA, Cowan CA, Garrison MM et al. Variation in inpatient diagnostic testing and management of bronchiolitis. *Pediatrics*. 2005; 115:878-884.
9. Ahmed FU, Karim E. Children at Risk of Developing Dehydration from Diarrhoea: A Case-control Study. *J Trop Pediatr*. 2002; 48:259-263.

APPENDIX

Sample of the DST output.

Features	Patient	Explanation case
Age	1.2	1.8
Birth	Vaginal	Vaginal
Smoking Mother	No	No
Hydration before treatment	Normal	Normal
O2 saturation before treatment	99.0	98.0
Retraction severity before treatment	None	Mod
Heart rate after treatment	129	129
Overall increase in work of breathing	None	None
Oxygen saturation under 92 after treatment	No (100.0)	No (99.0)
Respiratory rate over 60 after treatment	No (42)	No (38)
Temperature over 100.4 after treatment	No (98.0)	No (99.9)
Work of breathing after treatment	Same	Improved
Disposition		Admit

We suggest that this patient should be admitted to hospital.

In support of this prediction we have the Explanation Case that was older and had a better response to treatment but was still admitted to hospital.

However, it should be noted that the patient's lower heart rate after treatment and less severe retractions and higher O2 saturation before treatment in relation to the Explanation Case are features that go against our argument that the explanation case is healthier than the patient.

We have a reasonable confidence in our prediction

	Definitely Not	No	Maybe	Yes	Absolutely
Q1. Do you agree with the suggested course of action?	☐	☐	☐	☐	☐
Q2. Did you find the explanation case useful?	☐	☐	☐	☐	☐
Q3. Did you find the supporting dialog useful?	☐	☐	☐	☐	☐