Applying intention-based guidelines for critiquing

Robert-Jan Sips and Loes Braun and Nico Roos

Abstract. This paper investigates the combination of expert critiquing systems and formal medical guidelines. Medical guidelines might serve as a suitable basis for an expert critiquing system because of the ongoing acceptance of medical guidelines and the rise of both evidence-based practice and evidence-based guidelines. A prerequisite for a critiquing system based on medical guidelines is the ability to match the actions a physician performs in practice to actions prescribed by a guideline. Previous research has shown that this is quite difficult, due to the fact that computerized systems are unable to handle deviations from a guideline, which are common in the medical domain. Our solution to this problem is based on extracting the intention underlying a physician’s action and uses the intention as the basis for matching performed actions to prescribed actions. We propose an algorithm for an intention-based matching process and we evaluate the matching algorithm on 12 cases of hyperbilirubinemia in healthy term newborns.

1 Introduction

In the past three decades, various authors have expressed their worries about the state of medical practice [4, 8]. It is estimated that only 6% of all medical research after 1970 is scientifically sound [11]. Moreover, the boost of medical research in the past decades increased the difficulty of decision making for individual physicians, the “data overload has paradoxically led to a knowledge underload” [19]. To improve the quality of the medical practice and to support physicians in the process of decision making a number of tools have been proposed, including medical guidelines and expert critiquing systems [4].

Expert critiquing systems Expert critiquing systems were introduced to assist physicians in decision making, without forcing them to comply to a gold standard of care [16, 23]. Expert critiquing systems do this by providing critique on a physician’s decisions, rather than telling him exactly what to do. The first expert critiquing systems developed by Miller [16] relied heavily on user interaction, which resulted in rejection of his systems [9]. To avoid this problem, Van der Lei [23] proposed a system which gathers all its information from computer-stored medical records, critiquing a physician without relying on user interaction. The current developments, such as the introduction of an electronic patient record (EPR) in the Netherlands support this proposal.

A major problem with the critiquing systems proposed by Miller and Van der Lei is that these systems cannot cope with deviations from the underlying model. Moreover, the systems were not able to deal with the question why a physician was performing an action, which is essential to provide a grounded critique. To address both problems, Shahar et al. [20] and Advani et al. [1, 2] suggest to perform critiquing by assessing the compliance of a physician’s intentions with the intentions behind a medical guideline. They also stress that plan recognition is an indispensable prerequisite for the performance of critiquing.

Marcos et al. [14] claim that matching actions performed by a physician to those prescribed by the guideline is very difficult. Moreover, they claim that the direct use of intentions for critiquing, either for matching actions or for studying their appropriateness, is not possible. According to Marcos et al., this is caused by difficulties inherent to modeling and acquisition of these intentions: “The intentions of the guideline are not stated explicitly, which makes them very hard to model, and the intentions reported by experts almost always differ. It is not only a problem of vocabulary, but also a matter of differences in the degree of detail, abstraction level, etc.”.

In the paper we will show that recent developments in languages for describing medical guidelines combined with information treatment actions and medicines provided by insurance companies in The Netherlands enables us to overcome the objections state by Marcos.

Medical guidelines Medical guidelines were introduced to standardize medical practice. The use of medical guidelines have been shown to reduce practice variations [10], improve practice quality [21] and improve the cost efficiency of medical care [15]. Since the guidelines standardize medical practice, their introduction has faced great resistance from the medical field [4]. Physicians refer to standardized care as “cookbook medicine” [24], hampering the intellectual process of treatment and diagnosis. Moreover, the focus on medical-guideline development caused physicians to be “flooded with guidelines” [12].

Despite of these problems, guidelines are growing in number and acceptance around the world. Additionally, the quality of medical guidelines has drastically improved and is still improving because the old consensus-based guidelines (i.e., guidelines that reflect the opinion of a group of domain experts) are being replaced by evidence-based guidelines. The latter are guidelines based on scientific research, that thus reflect the latest of medical knowledge in the field [24].

To improve medical guidelines [15] and to make medical guidelines computer understandable, research has been conducted on the formalization of medical guidelines, using formal modeling languages [18]. The most-developed examples [22] of these formal languages are ASBRU, EON, GLIF, GUIDE, PRODIGY and PROFORMA. We evaluated these languages to determine to what extent they can be used for critiquing systems.

1 Department of Computer Science, Universiteit Maastricht, P.O.Box 616, 6200 MD Maastricht. Email: baun@cs.unimaas.nl
2 Department of Computer Science, Universiteit Maastricht, P.O.Box 616, 6200 MD Maastricht. Email: roos@cs.unimaas.nl

Originally denoted as Proforma [22].
For intention base critiquing, it is important that the intentions are formulated in a computer interpretable manner, i.e., in a formal manner. ASBRU, EON, GUIDE and PROFORMA are the only languages which specify intentions formally, from which ASBRU has the most extensive intention-modeling capacity [17]. Moreover, the ASBRU guideline model is the simplest, containing only one generic plan object. In contrast to the other languages, ASBRU and PROFORMA, can model sequential and parallel execution, without introducing a new modeling concept. Therefore, we have chosen to investigate a critiquing system that uses a medical guideline described in ASBRU.

Paper outline In the next section we describes our approach of combining medical guidelines and critiquing systems and section 3 presents an experimental evaluation of our approach. Section 4 concludes the paper.

2 Intention-based treatment identification

Intentions Medical actions vary from medicine prescriptions to blood tests and from examining x-rays to surgical interventions. All of these actions can be used in a different context within different treatments. Measuring one’s blood type (referred to in literature as an ABO test [3]) could be used in, for instance, a guideline containing a blood transfusion to determine the applicability of the blood for transfusion. However, in a guideline for Jaundice in newborns, the guideline can prescribe the same action to determine the likeliness of a blood-group antagonism in the newborn, i.e., the likeliness that the newborn has antiserum to its own blood [6]. Because the intention of an action can change based on the context in which the action is performed, it is hard to model the possible intentions of an action.

When we re-examine the example above, we see that in all cases the ABO test is used to determine the patients’ blood group. This is a context independent intention of this action. Based on these observations, we define two types of intentions:

- **High-level intentions** Specifying an intention of an action in a specific context, e.g., “determine the possibility of blood group antagonism”.
- **Low-level intentions** Specifying a context independent intention of an action, e.g., “determine blood group”.

The high-level intentions relate to the intentions of medical guidelines and low-level intentions relate to a physician’s actions and to actions in medical guidelines. In the following paragraphs we will explain how we can exploit the relation between high-level and low-level intentions and medical guidelines.

Medical guidelines Based on the physician’s actions reported in the EPR, the critiquing system must determine to what extent the guideline has been followed by the physician in order to provide adequate critique. Hence starting form a general high-level intention such as “diagnose and treat hyperbilirubinemia”, the critiquing system must search through the possible execution of the guideline. We interpreted this as a refinement of intentions and we will illustrate the refinement process using parts of the guideline for diagnosis and treatment of hyperbilirubinemia, which was developed in the Asgaard project [15]. The critiquing system starts with searching for a plan to refine the high-level intention “diagnosis and treat hyperbilirubinemia”.

| plan Diagnostics-and-treatment-hyperbilirubinemia
| intentions avoid intermediate-state: (bilirubin = transfusion)
| conditions complete-condition:
| or:
| (jaundice-clinically-significant = no)
| explanation "Exit the guideline to individualized clinical evaluation, including assessment of jaundice in light of prematurity."
| (age = 1 day)
| explanation "Exit the guideline to individualized clinical evaluation, including the assessment of jaundice and non-isoimmune hemolytic disease."
| (pathologic-reason = yes)
| plan-body type = sequentially
| wait-for none
| ask term-child
| ask age-child
| Diagnostics-hyperbilirubinemia
| Treatment-hyperbilirubinemia

On inspection of the plan, the critiquing system will subsequently check the EPR whether the condition “jaundice-clinically-significant = no” is among the preliminary reported facts. If this is not the case, it searches for a plan that realizes the high-level intention “is-known-parameter (jaundice-clinically-significant)”. The following plan realizes this intention.

| plan Jaundice-determination
| intentions achieve overall-state: is-known-parameter (jaundice-clinically-significant) in NOW
| plan-body type = sequentially
| wait-for all
| do type = any-order
| wait-for one
| Blanching-skin-with-digital-pressure-test
| Icterometer-test
| Transcutaneous-jaundice-meter-test
| Determine-extent-cephalocaud-progression
| ask jaundice-clinically-significant

After having refined the high-level intention behind the conditions, the critiquing system continues with the plan body. When processing the plan body, it also treats the sub-plans as high-level intentions. Hence, the high-level intention “Treatment-hyperbilirubinemia” is realized by the plan:

| plan Treatment-hyperbilirubinemia
| intentions avoid intermediate-state: (bilirubin = transfusion)
| achieve overall-state: (bilirubin = observation)
| plan-body type = parallel
| wait-for one
| wait-for
| (or Regular-treatments Exchange-transfusion)
| Regular-treatments
| on-abort Exchange-transfusion
| Exchange-transfusion
| cyclical-plan
| do type = sequentially

...
Physician’s actions and lower-level intentions \hspace{1em} To explain the trace of actions performed by a physician and reported in the electronic patient record, we have to match these actions with the actions prescribed by one of the execution sequences of a guideline.

Directly matching executed actions with prescribed actions is hard. As was pointed out by Woolf et al. [24], medical practice variations are mostly caused by factors as the treating physician, the hospital or the geographical location. For instance, when a physician prescribes a short working blood-glucose-lowering substance, e.g., Velosulin, the intention of lowering the blood-glucose level could just as well be served by prescribing Humalog, or Novorapid. In other words, one physician may prefer Velosulin, because it is available in the local drug store, or because he has good experiences with it, while the other, who is, for example, located 20 miles south, may prefer Novorapid, for the same reasons. Hence, because of these seemingly arbitrary factors influencing a physician acting, there is no point in trying to match guideline-prescribed and observed actions. Therefore, we replace both the guideline-prescribed actions and the observed action by low-level intentions.

In the Netherlands, some types of actions are thoroughly documented and restricted by the health care insurance companies. For example, the types and brand names of medicines which can be prescribed in the Netherlands are documented in the Pharmaco-therapeutisch kompas (pharmacotherapeutic compass) [7]. The pharmacotherapeutic compass groups medicines with similar effects into pharmaceutical groups. For instance, Velosulin, Humalog and Novorapid are in the same pharmaceutical group, but differ in brand names.

In order to reason about the low-level intentions of a medicine prescription, we translate medicine prescriptions to pharmaceutical groups. So the system can reason about, e.g., the prescription of a short working blood-glucose lowering substance, instead of the prescription of Velosulin. We believe that this translates more naturally to a physicians thinking. In a similar way, we have used the Merck-manual [3] for translating treatment-actions to low-level intentions.

The above mentioned medical literature is almost complete with respect to the set of action (medical interventions). Moreover, there is a trend to standardize the terms used to describe actions reported in the EPR. Hence, it is possible to build a translation table specifying all possible intentions of each action mentioned in the guideline or reported in the EPR.

Matching \hspace{1em} To identify the execution sequence of the guideline a physician is following, we define a measure of the distance between an action sequences prescribed by an execution of the guidelines and the action sequences reported in the EPR.

The distance is defined as (1) the number of actions in de EPR that are matched by the execution sequence of the guidelines minus (2) the number of unmatched actions in the execution sequence that occur before the last matched action in the execution sequence.

Note that this distance measure does not take into account the order of the actions prescribed by the guideline. This is a deliberate choice which has been made in order to reduces the effect of variations in the order in which actions are reported in the EPR. By minimizing the unmatched actions that occur before the last matched action, we try to avoid matching the reported actions with a wrong execution sequence.

A match between an action prescribed by a guideline and a physician’s action is preformed using the above mention translation table and is defined as follows.

A match is an action performed by the physician with a low-level intention that is contained in either (1) the low-level intentions of one of the guideline actions, (2) one of the guideline’s action effects or plan effects, (3) a guideline plan precondition.

In the following algorithm for matching the actions performed by a physician to those prescribed by a guideline, we assume that we know the most general high-level intention of the physician; e.g., "treat hyperbilirubinemia":

1. Let $e = \langle i_1, ..., i_n \rangle$ be an execution sequence of low-level intentions realizing a high-level intention $h$. We can then define $E(h)$ to be all legal execution sequences for a high-level intention $h$. $h$ is the most general high-level intention of the guideline.

2. Let $A = \langle a_1, ..., a_m \rangle$ denote the physician’s actions. Then $t(A) = \langle t_1, ..., t_m \rangle$ is the corresponding low-level trace for the physician’s actions.

We can now define the minimal execution sequence $e$ that explains the maximal number of actions in $A$. Note that no two $t_i$ in $t(A)$ can be matched to the same $i$ in $e$.

3. Let $M \subseteq \{1, ..., m\} \times \{1, ..., n\}$ be a mapping from the indexes of a trace $t = \langle t_1, ..., t_m \rangle$ to the indexes of the execution sequence $e = \langle i_1, ..., i_n \rangle$ satisfying the following four conditions:

   - $(i, j) \in M$ iff $t_i = i_j$;
   - for every $(i, j) \in M$ and $(x, y) \in M^2$: $i = x$ iff $j = y$;
   - if no $j$: $(i, j) \in M$, then $i_j \not\in \{y \mid \exists x : (x, y) \in M\}$;
   - there is no mapping $M'$ satisfying the first three items above such that:
     $\{i \mid (i, j) \in M'\} = |\{i \mid (i, j) \in M\}|$ and
     $\max(\{j \mid (i, j) \in M\}) < \max(\{j \mid (i, j) \in M'\})$.

The first item guarantees that there are only mappings between identical intentions. The second item guarantees that there are only one-to-one mappings. The third item guarantees that all possible intentions of a trace are mapped to an intention in an execution. Finally the last item minimizes the non-matched gaps in the execution.

4. Let $k = \max\{j \mid (i, j) \in M\}$. Then the measure of the match between a trace $t$ and an execution $e$ is defined as:

   \[ \text{match}(t, e) = |M| - |\{j \mid j < k, (i, j) \in M\}| \]

5. The best match for the physician’s actions $A$ is the execution $e \in E(h)$ for which $\text{match}(t(A), e)$ is maximal.
3 Experiments

To test our approach described in the previous chapter, we implemented a prototype. This prototype uses a guideline formulated in ASBRU and a translation table for mapping actions to sets of low-level intentions. Our prototype is capable of determining all possible executions through an ASBRU modeled protocol automatically, without taking temporal constraints into account. In our experiments we have used an ASBRU-modeled guideline, constructed in the Asgaard project [15]. This guideline models the diagnosis and treatment of hyperbilirubinemia in healthy newborns based on the Jaundice guideline of the American Association for Pediatrics (AAP), which is intended for the management of Jaundice in healthy term (defined 37 completed weeks of gestation) newborns [13]. The translation table has been constructed using the pharmacotherapeutic compass [7] and the Merck-manual [3].

Hyperbilirubinemia or Jaundice is a common disease in newborn babies and caused by a elevated blood-bilirubin level. In many cases Jaundice disappears without treatment, but in some cases phototherapy (treatment with UV radiation), or even blood-transfusion is needed to reduce the high bilirubin level. Jaundice can be an indicator for a number of serious disorders, including hemolytic diseases, liver disorders, fetal-maternal blood group incompatibility, etc. For a full overview we refer to the Merck-manual of treatment and diagnosis [3], page 2156.

Marcos et al. provided us with the test data used in their critiquing experiment [14]. Marcos et al. obtained their data from a pediatrician addressing patient cases. Since the pediatrician consulted by Marcos et al. received the guideline prior to addressing the patient cases, we asked Twan Mulder MD Phd, who works as a pediatrician / neonatologist at Maastricht University Hospital to address patient cases without prior knowledge of the guideline. In this way, we gathered solutions to 12 cases. The pediatrician consulted by Marcos et al. provided solutions to case 1-8. Case 7 was discarded, because the pediatrician had trouble with the used terminology. T.Mulder provided solutions to case 8-12.

Methods Each solutions of the pediatricians has been used to generate a set of test executions. A test execution is a prefix of the physician solution, always starting at the first action. These test executions simulate the step by step updating of the EPR by a physician.

In the first experiment we measured the performance of our algorithm on all test executions based on the normal execution order. With this test we wish to determine the following.

- The overall performance of our algorithm (which percentage of a sequence is correctly classified).
- Whether there is a significant difference in performance on the solutions of a physician who was familiar with the guideline used and one who is not.

In the second experiment we measured the performance of our algorithm on the solutions in a different order. This was done to obtain a measure of the performance on a real patient record, where the physician enters the information on consultation basis and where it is not guaranteed that the actions are entered in the order in which they are performed. To do so, we used two execution orders, reason by encounter (RBE) and backwards execution. In reason by encounter, we assumed performs 3 action per consult and these 3 actions need not be entered into the electronic patient record in the order in which they were performed. Backwards execution was used to determine the performance of our algorithm on actions in the completely opposite order, i.e., backwards execution. Based on these experiments, we statistically tested the differences between these three execution orders to determine the dependence of our algorithm on the correct sequence of actions.

Results First of all we determined the overall performance of our algorithm. For each of the 12 sequences provided by the physicians, we generated new sequences by considering all prefixes of the complete sequence. The prefixes represent the available data after every step in the treatment process. Figures 1 and 2 show the average results of matching the sequences with the guideline.

![Figure 1](image1.png)

**Figure 1.** Average results of matching Marcos’ actions to the guideline. $avg$ denotes sample mean and $avg \pm s$ denote the sample variance.

![Figure 2](image2.png)

**Figure 2.** Average results of matching Mulder’s actions to the guideline.

On inspection of the experimental results when using the normal execution order, we see that, on average, 69.4% of the actions in a test sequence are matched correctly. When we have a closer look at the misplaced action types, we see that 19 out of the 27 misplaced action types cannot be matched due to incompleteness of the guideline. The
The ASBRU guideline is incomplete with respect to the specification of treatment for several causes of hyperbilirubinemia. The following plan gives an illustration.

plan Check-for-rapid-TSB-increase
  intentions
  achieve overall-state:
  (and is-known-variable(possibility-of-G6PD)
   is-known-variable (possibility-of-hemolytic-disease))
  conditions
  filter-precondition:
  (and (TSB - decrease = no) in NOW
   (TSB - change > 0.5) in NOW)
  plan-body type = sequentially
  wait-for all
  possibility-of-hemolytic-disease <- yes if (age = day2) then
  possibility-of-G6PD <- yes
  Exit-possibility-of-G6PD
  else
  possibility-of-G6PD <- no
  Exit-possibility-of-hemolytic-disease

When we discard the actions which cannot be matched, we observe an average performance of 80.1%.

We can also see that the average performance on short test sequences is worse than on long sequences. This can be explained by looking at the actions reported by the expert. It turns out that in most cases the first action of the pediatrician is "Measure TSB-value", which (according to the reported intention), should be matched to "ask TSB-value" in "Treatment Hyperbilirubinemia". However, our algorithm is only able to do this in sequences where the pediatrician has prescribed treatment already. In other cases, there is an execution which misses less guideline steps (that is, an execution in which the action is matched to "ask TSB-value" in "Check for Jaundice after 3 Weeks"). Looking at the Merck manual and the Diagnostic Compass, we find that the TSB value, (Total Serum Bilirubinemia) is the criterion for the presence of hyperbilirubinemia. The Merck manual defines hyperbilirubinemia as:

"A serum bilirubin concentration > 10 mg/dL in preterm newborns or > 15 mg/dL in full-term newborns"

Thus the lack of an initial measurement of TSB indicates a gap in the guideline (or in its ASBRU counterpart). This is also confirmed by the pediatrician’s intention, to determine the level of hyperbilirubinemia.

When we discard the action "measure TSB-value", which cannot be matched correctly due to a gap in the guideline and the actions which have no match in the guideline, we observe that, on average, our algorithm matches 95.8% of the actions in a sequence to the correct action in the guideline.

Based on the results on the normal execution sequences, we investigated whether there is a significant difference in the performance of our algorithm on solutions of a pediatrician familiar with the guideline (case 1 - 8) and on solutions of a pediatrician unfamiliar with the guideline (case 8 - 12). Because it is clear that the performance of both physicians relative to case size would give biased results4, we chose the absolute number of misplaced actions per sequence as comparison statistic. On average, our algorithm misplaced 2.62 actions per sequence for Mulder and 1.59 actions per sequence for the pediatrician consulted by Marcos et al. This indicates a better performance on the pediatrician consulted by Marcos et al. To test whether this difference is significant, we conducted an approximate randomization test, as suggested by Cohen5 in similar cases. This test showed no significant difference on a 97.5% level of confidence.

Finally, we were interested in the question whether there is a difference in the performance of our algorithm on normal, RBE and backwards executions. Our results show no significant difference on a 97.5% level of confidence.

4 For example, if the 2nd action of a sequence is misplaced, the performance is lowered by 50%. However, when the 3rd action is misplaced, the performance is lowered by 33%. In comparison however, every misplaced action is equally important.

4 Conclusion

In this paper we presented a framework for intention-based matching of a physicians actions reported in an electronic patient record (EPR), to those prescribed by a medical guideline. Our framework was implemented and tested using a guideline for the diagnosis and treatment of hyperbilirubinemia and using treatment data provided by two pediatrician for a set of test cases. One of pediatrician was familiar with the guideline that we used and one who was not. In section 3, we concluded that on average 69.4% of a treatment plan proposed by a physician is matched correctly to the guideline. After correcting for the the results for the incompleteness of the guideline used, on average 80.1% of a treatment plan proposed by a physician is matched correctly to the guideline.

We observed that the matching results are significantly better for long sequences of treatment actions preformed by physician. The reason for this is that the first action the pediatricians reported could not be matched onto the guideline correctly because of an error in the guideline. When we also discard this action, on average 95.8% of a treatment plan proposed by a physician is matched correctly to the guideline. This result shows that the physicians follow the guidelines rather well. Moreover, it provides a solid base for a critiquing system. When a physician is skipping expected actions or performing unexpected actions, there is a reasonable chance that the feedback given by the critiquing system is relevant.

Based on the experimental results, we also investigated whether there is a significant difference in the performance of our algorithm on solutions of a pediatrician familiar with the guideline and on solutions of a pediatrician unfamiliar with the guideline. Though the results indicate a better performance on the pediatrician familiar with the guideline, a statistical test showed that the difference is not significant.

Finally, we investigated the influence of entering data into the electronic patient record (EPR) on a consultation basis where there is no guarantee that the action are entered in the order in which they are performed. Though this resulted in a decrease of the performance, a statistical test showed that the difference was not significant.

The measure that we used to determine the distance between the sequence of actions prescribed by the guidelines and the performed actions reported in the EPR does not explicitly take deviations with respect to the order prescribed by guideline into account. This has been a deliberate choice made to increase the robustness of the matching process with respect to variations in the order of performed actions. As a side effect, the current matching process is sensitive to errors in the guideline. Future research will therefore focus on developing better measures that can cope with errors in the guideline without increasing the sensitivity to variations in the order of performed actions and, secondly, can use the temporal data provided by an ASBRU guideline. Moreover, future research will also address more complex situations, such as the most general intention being unknown and the treatment of more than one disorder at the same time.
REFERENCES


