DELIVERABLE SUMMARY SHEET

Project Number: ***IST–2001-33049***

Project Acronym: **PROTOCURE**

Title: ***Improving medical protocols by formal methods***

Deliverable N°: **D4**

Due date: **30/9/2002**

Delivery Date: **9/10/2002**

Short Description:

This document describes and comments on the results achieved in the work directed to formally verify the two reference guidelines selected in the project (WP4). It lists all the results we have obtained, which can be summarised as follows.

We have started with the list of desirable/required protocol properties that we had included in the deliverable D3, issued from the workpackage on properties of Asbru protocols (WP2). Of these properties, only a subset was considered for formal verification. The main reason for that was the tight time schedule of this assessment project.

Thus, we have formally verified various properties, with different types of results. In some cases the properties have been confirmed, which means that the formalised version of the protocol complies with them. However, in other cases formal verification was problematic, i.e. it was not possible to complete the proofs unless we added additional assumptions describing the necessary conditions that make the property true. In both cases we can say that formal verification is beneficial and has advantages over other approaches to improve the quality of protocols.

Partners owning: *all*

Partners contributed: Universität Augsburg, Vrije Universiteit Amsterdam, University of Aberdeen

Made available to: Public

Verification of Properties on the Reference Protocols
Version 1.0

October 8, 2002

Contents

1 Introduction

This document describes and comments on the results achieved in the work directed to formally verify the two reference guidelines selected in the project (WP4). The first guideline that was considered concerned jaundice in the newborn (also called icterus neonatorum). The second one concerned the management of diabetes mellitus type 2 (also called non-insulin dependent diabetes mellitus). These two guidelines share some features, for example in the sense that time plays an important role in both guidelines. In both cases medical management can be interpreted as hierarchical planning.

However, the guidelines are also different: they deal with different medical conditions for which objectives differ, and also the structure of the guidelines is dissimilar. As the jaundice guideline is the one investigated initially, this is also the one where most of the experience in employing a verification tool such as KIV was built up. This experience made clear that a similar activity would be feasible for the diabetes type 2 guideline as well. As arguably one of the most difficult issues of this work is to ensure that flaws that are discovered using our methods are also medically important, or at least relevant, the diabetes type 2 guideline verification work was slightly refocussed in order to acquire some insight into these issues as well. Obviously, obtaining more insight into how the gap between medicine and formal methods can be bridged is a necessary requirement for future progress.

Concerning verification tasks, this document lists all the results we have obtained, which can be summarised as follows. We have started with the list of desirable/required protocol properties that we had included in the deliverable D3, issued from the workpackage on properties of Asbru protocols–WP2. Of these properties, only a subset was considered for formal verification. The main reason for that was the tight time schedule of this assessment project. Thus, we have formally verified various properties, with different types of results. In some cases the properties have been confirmed, which means that the formalised version of the protocol complies with them. However, in other cases formal verification was problematic, i.e. it was not possible to complete the proofs unless we added additional assumptions describing the necessary conditions that make the property true. In both cases we can say that formal verification is beneficial and has advantages over other approaches to improve the quality of protocols. We will illustrate this in Section 7 with a sort of platonic dialog between a medical expert and a verifier/formal method specialist.

The document is structured as follows. Section 2 summarises the results of our efforts in WP4. Details on the verification of Jaundice and Diabetes properties are given in Sections 3 and 4. Examples of proofs can be found in the Appendix part. Appendix A contains a rather simple proof and Appendix B describes our effort to prove a more complicated property. The latter has been organized as a self-contained document. Besides the contributions to the quality improvement of medical guidelines, our efforts gave rise to a number of spin-offs (see Section 5) and ideas for future improvements (see Section 6). Finally section 7 closes the document with a discussion of the obtained results in the form of a dialog between a medical expert and a verifier/formal method specialist.

2 Summary of results

We have considered two medical guidelines, Jaundice and Diabetes. For Jaundice, our medical experts have proposed ten properties which the guideline should adhere to. These properties are listed in Section 3. From this list, we have chosen four properties which were considered the most relevant ones from the medical point of view, and additionally challenging for fornal verification. We were able to confirm two of the properties: they are definitely satisfied by the Jaundice guideline (see Sections 3.2 and 3.3). For the two other properties, we have found counter examples describing the circumstances under which they are not fulfilled (see Sections 3.1.2 and 3.4). For Diabetes, we started with a list of 22 properties (see Section 4). Since many of these properties, although medically relevant, did not require the use of formal verification techniques, we have explored the utilisation of a model of the disease as an additional resource for the definition of protocol properties.

In summary, we can say that formal methods have been beneficial to the improvement of the quality of our medical guidelines in several ways:

1. *Formal verification validates Asbru model*

   Constructing an Asbru model for a guideline in natural language is not straightforward, because details need to be added. At first, an Asbru model contains minor faults due to misinterpretation of the guideline text. Formal verification can be used to validate the Asbru model and to detect

these errors.

2. *Formal verification confirms properties or reveals flaws*

   For example, if a medical guideline is developed with certain intentions in mind, formal verification can be used to analyse the guideline and to determine if the intentions are achieved. If not, counter examples can be obtained and hence the guideline can be improved so that they are taken into account.

   As another example, we have analysed one of the criteria (or indicators) established by a medical organisation for the management of Jaundice (see Section 3.4 and Appendix B), and discovered that the original guideline does not satisfy the indicator under all circumstances. These circumstances were illustrated by a concrete counter example. In addition to this, the proof attempts gave rise to three assumptions which must be fulfilled for the indicator to hold. These assumptions can be used by medical experts to either restrict the indicator or to improve the original guideline.

We have shown that formal methods can help to improve the quality of medical guidelines. However, formal verification was harder than expected and finally we have only been able to verify a few properties, mainly because tool support had to be developed during in parallel to the verification tasks. Besides, more efficient formal verification requires a more elaborate tool support. Our proof attempts have served to better determine a number of requirements that would be necessary for an efficient verification. These requirements have been listed in Sections 5 and 6. With these improvements, we are confident that verification will be much more efficient in the future.

3 Verification of Jaundice

In this section we revisit the protocol properties that we considered promising for verification purposes from a medical point of view. These properties were part of deliverable D3, about desirable/required protocol properties. We will describe the properties again, together with some indications of the means (formal or others) we have used to verify them.

More precisely, for each property we will describe:

- Informal explanation of the property
- Role of the property
- Considered for formal verification?
- If no, why?
- If yes, formalisation
- Summary of proof
- Results

In this section we first revisit the jaundice properties mentioned in deliverable D3, in this way, and then take a closer look to the verification of some of them: the intentions of plans Phototherapy-intensive and Phototherapy-normal-prescription, and the MAJIC indicator #7.

3.1 Properties

3.1.1 Protocol-dependent properties

Correctness of intentions This general property aims at ensuring that the intention of a plan follows from both the intentions of its subplans and the way these are composed.

Role Checking that plan intentions are fulfilled.

Considered? Yes (see examples in sections 3.2 and 3.3).

Correctness of phototherapy treatment Ensure that phototherapy treatment is applied properly, with a maximum surface of the skin exposed to the light.

Role Treatment effectiveness.

Considered? No.

Why? Not applicable. There is no mention to this aspect among the recommendations of the protocol. However, the protocol describes the factors that can improve the efficiency of phototherapy (including the surface area that is exposed to the lights), within an appendix.

Safety of phototherapy treatment Ensure that phototherapy treatment is not applied for a time longer than 10 days.

Role Safety consideration.

Considered? No.

Why? After review by a jaundice expert, this property was not considered valid.

Sequences of actions for jaundice management This property involves checking the actions performed by a physician to solve a particular case against the recommendations of the protocol for that case. It can be seen from two points of view: first, we can determine whether the doctor complies with the standard recommendations, and second, we can check whether the protocol reflects the physician's actions, which are in this case considered as standard.

Role Critiquing of physician's actions or validation of the protocol.

Considered? No.

Why? Necessary KIV support was not yet implemented.

Correctness of actions after phototherapy treatment Ensure that, once phototherapy has been applied, bilirubin measurements are performed and not blanching skin tests.

Role Action correctness.

Considered? No.

Why? Trivially satisfied in the original protocol: see Evaluation and Treatment sections of the protocol (blanching skin test can only applied in the Evaluation section, before the baby can go under phototherapy).

Correctness of actions before and during phototherapy treatment Ensure that the baby is hydrated before and during phototherapy.

Role Action correctness.

Considered? No.

Why? Not applicable. Although the protocol cites in its appendix hydration as an aspect that should contribute the efficacy of phototherapy, there is no mention to it in the recommendations of the protocol.

Correctness/safety properties in terms of indicators This general property aims at verifying that the application of the protocol results in actions that comply with indicators defined either in the protocol itself or by external sources. For this purpose, it is possible to use the clinical indicators for jaundice treatment that have been defined by the MAJIC (Making Advances against Jaundice in Infant Care) Steering Committee.

Role Checking that indicators are satisfied.

Considered? Yes (see example in section 3.4).

3.1.2 Other properties

Termination Does execution of the formal model of the guideline eventually terminate?

Role Basic correctness property. Useful for validating formal model.

Considered? Yes.

Formalisation To verify that an Asbru plan terminates, the following proof obligation needs to be verified.

$$\langle \textit{plan} \rangle \rightarrow (\Box \langle \textit{plan-filter} \rangle \rightarrow \Diamond \textbf{laststep})$$

If a plan is considered and its filter condition is satisfied, then eventually the last step of execution is reached.

Summary of proof The property has been successfully verified for a number of subplans of Jaundice. Modular proofs for plans referring to subplans have been constructed. Finally an attempt to verify termination for Regular-treatments has been tried. However, this plan does not terminate in every case.

Because proofs are very similar, it was not very challenging to prove termination for all subplans of Jaundice.

Results The following subplans terminate:

- Prescribe-intensive-phototherapy
- Phototherapy-intensive
- Prescribe-normal-phototherapy
- Phototherapy-normal-prescription
- Prescribe-normal-phototherapy-manual
- Prescribe-observation-manual
- Phototherapy-normal-recommendation
- Prescribe-observation
- Observation
- Breastfeeding-manual
- Breastfeeding-with-formula-manual
- Formula-only-manual
- Feeding-alternatives

However, the plan Regular-treatments does not terminate in every case.

3.2 Intention of Phototherapy-intensive plan

Informal explanation While executing Phototherapy-intensive, there should be a decrease in level of TSB of at least 10mmol/l within 4 to 6 hours.

Role Checking that plan intentions are fulfilled.

Considered? Yes.

Formalisation The above property has been modelled in Asbru as follows:

MAINTAIN INTERMEDIATE STATE
($TSB\text{-}decrease = \text{yes}$) $([4h, -][-, 6h][-, -], self)$
AND ($TSB\text{-}change \geq 10$) $([4h, -][-, 6h][-, -], self)$.

This intention is translated to the following temporal formula in KIV:

```
/* intention */
□  laststep
  ∨    pti-state ≠ activated ∧ pti-state′ = activated
    →          /* time annotation */
                 4 ≤ time − pti-acttime .value
               ∧ time − pti-acttime .value ≤ 6
             → /* property */
                 get-decrease(pd .tsb, pti-acttime .value)
               ∧ get-change(pd .tsb, pti-acttime .value) ≥ 10
        unless pti-state′ ≠ activated
```

For a detailed explanation of this formula, see Section 5.1.2.

Summary of proof The property has been successfully proven in KIV. The proof is fully automatic. A summary of the proof can be found in Appendix A.

Results It has been formally proven that the intention is satisfied by plan Phototherapy-intensive. During early proof attempts, a number of errors have been uncovered in the formalisation of intentions. This led to an improved formalisation of the intention (see Section 5.1.2), and to better insight into intention formalisation patterns. Thanks to this, the semantics of Asbru intentions is now more precise.

3.3 Intention of Phototherapy-normal-prescription plan

Informal explanation While Phototherapy-normal-prescription is applied, there should always be a decrease in TSB level.

Role Checking that plan intentions are fulfilled.

Considered? Yes.

Formalisation The above property has been modelled in Asbru as follows:

MAINTAIN INTERMEDIATE STATE *TSB-decrease* = yes

Compared to the intention of Section 3.2, this intention does not contain any time annotation. Therefore the translation to a temporal formula in KIV has been much simpler.

$$
\begin{array}{l}
\Box \quad \textbf{laststep} \\
\quad \vee \quad \textit{ptnp-state} \neq \text{activated} \wedge \textit{ptnp-state}' = \text{activated} \\
\quad\quad \rightarrow \quad (\text{get-decrease}(\textit{pd}\ .\text{tsb}, \textit{ptnp-acttime}\ .\text{value}) \\
\quad\quad\quad \textbf{unless}\ \textit{ptnp-state}' \neq \text{activated})
\end{array}
$$

Always, either we are in the last step or, if Phototherapy-normal-prescription is activated, there will be a decrease in TSB level, if we compare it to the level at activation time (*ptnp-acttime*) of this plan.

Summary of proof The property has been successfully proven in KIV. The proof is fully automatic. The complexity of the proof is comparable to the example proof of Appendix A.

Results It has been formally proven that the intention is satisfied by plan Phototherapy-normal-prescription.

3.4 MAJIC indicator #7

The MAJIC indicators that appear in [3] have been refined by the same organisation into different medical review criteria for the evaluation and treatment of jaundice. This includes a set of 11 criteria together with their inclusion conditions. Among these, we have selected indicator #7, which is stated as follows:

INCLUSIONS If any phototherapy initiated

CRITERIA No more than one serum bilirubin level drawn after phototherapy is discontinued

The rest of criteria were discarded mainly for two reasons. First, some of them refer to to terms/circumstances that neither are defined nor appear in the protocol, like "effective phototherapy" or "c-section delivery discharged prior to 72 hours of age". Second, many of them roughly relate the age of the baby with the phototherapy treatment depending on the TSB[1] level. This makes the verification of such indicators less challenging, because it can be tackled by manually checking Table 2 of the original protocol, which contains this information.

Below we present a formulation and a sketch of the verification of MAJIC indicator #7. For more details on the proof, see appendix B.

Informal explanation Only one more TSB measurement is performed after phototherapy is discontinued.

Role Minimizing risky measurements, cost effectiveness

Considered? Yes.

Formalisation The indicator has been translated into the following temporal formula.

$$
\begin{array}{l}
\textit{/* indicator */} \\
\Box \quad \textbf{laststep} \\
\quad \vee \quad pd \text{ .under-phototherapy} \wedge \neg\ pd' \text{ .under-phototherapy} \\
\qquad \wedge\ pd \text{ .tsb} = TSB_0 \\
\qquad \rightarrow \Box\ pd \text{ .tsb} = TSB_1 \wedge TSB_1 \neq TSB_0 \rightarrow \Box\ pd \text{ .tsb} = TSB_1
\end{array}
$$

Section B.3.3 explains the translation.

Summary of proof We attempted to prove the indicator for plan Regular-treatments. Several proof attempts have been made which uncovered a number of errors in the formalisation of Jaundice. These errors were used to improve the formalisation pattern for Asbru plans in KIV (see Section 5.1.1). In addition to this, a counter example for the indicator has been generated. Thus the indicator is not satisfied by the Jaundice guideline. The proof attempts gave rise to three assumptions which should be strong enough for the indicator to hold. The final proof is not yet completed, because tool support needs to be improved first. For more details see Appendix B.

[1] Total serum bilirubin.

Results Property does not hold.

A counter example is to apply phototherapy once and then to do observation for more than 24 hours allowing Treatment-hyperbilirubinemia to measure TSB twice.

Under the following assumptions the indicator should be satisfied:

1. If phototherapy is discontinued, TSB (reading) is observation.
2. Observation will run for less than 24 hours.
3. After phototherapy and observation, TSB (reading) will still be observation.

Several errors in the formalisation of Jaundice in KIV have been uncovered.

4 Verification of Diabetes

Like in the previous section, here we revisit the diabetes properties mentioned in deliverable D3. Besides, we present other diabetes properties that our medical partner considered important, although they were not included in D3. In all cases we informally state the property and give some indications of the means that we have used to verify it.

In the case of diabetes protocol, as we will see, many of the properties turned out to be either not applicable or not interesting for formal verification purposes. The former category includes e.g. properties that are related to features of the disease that are not covered in our protocol, like e.g. "individuals unaware of hypoglycemia". As for the latter, most of the properties could be checked by analysing if the original or the Asbru version of the protocol verified them or not. However, it is important to note that in two cases we had to resort to the Asbru protocol, which includes more precise information in some parts –information which was introduced with the help of our medical partner to complement/clarify the obscure points in the original document. Finally, the formal verification of the remaining properties –basically, plan intentions– was not considered either due to the tight time schedule of this assessment project.

The issue of property definition is difficult, since we have experienced that properties that are interesting from the medical point of view could be difficult or even impossible to verify, depending on the way formalisation has been tackled. This has led us to explore alternative ways of defining protocol properties, including the utilisation of a model of the disease mechanisms (in this case, diabetes). There are many advantages in the use of such a model, like e.g. its utilisation as background knowledge in the formal verification process.

In the first subsection we will describe, for each of the diabetes properties:

- Informal explanation of the property
- Role of the property
- Considered for formal verification?
- If no, why?

Then, in the second subsection we will present our ideas about the utilisation of a model of the disease for both the definition of more relevant protocol properties and their verification.

4.1 Properties

4.1.1 Protocol-dependent properties in deliverable D3

Correctness of doses in insulin treatment When using insulin in 2 doses per day, ensure that the morning and evening doses are distributed according to the ratio 2/3-1/3.

Role Treatment efectiveness.

Considered? No.

Why? Trivially (without formal verification) violated in the original protocol: see transition from 1 to 2 doses per day in 'Insulin plus oral antidiabetics' treatment.

Safety of dose increase in treatment phase Ensure that drug/insulin doses are not increased too quickly unless the patient is in a critical situation. More precisely, the dose should be increased by the smallest amount in intervals of at least 2 to 4 weeks, unless the blood glucose is above 15 mmol/l.

Role Safety consideration.

Considered? No.

Why? Timing is trivially (without formal verification) violated in the Asbru protocol: e.g. insulin dose increase is performed every 7-10 days, according to our expert (this information was lacking in the original protocol).

Periodicity of controls in patients who have suffered a foot ulcer Ensure that these patients have their feet inspected at least every 3 months.

Role Action correctness.

Considered? No.

Why? Not applicable. There is no mention to these particular patients in the protocol.

Safety of management of patients with hypertension Ensure that individuals who are unaware of hypoglycemia do not get beta-blockers for the treatment of hypertension.

Role Safety consideration.

Considered? No.

Why? Not applicable. There is no mention to this feature (hypoglycemia unawareness) in the protocol.

4.1.2 Additional protocol-dependent properties

Correctness of intentions This general property aims at ensuring that the intention of a plan follows from both the intentions of its subplans and the way these are composed.

Role Checking that plan intentions are fulfilled.

Considered? No.

Why? Time constraints.

Correctness of actions in diagnosis phase Verify that the oral glucose tolerance test (OGTT) is not applied in GP practice.

Role Action correctness

Considered? No.

Why? Not applicable. There is no mention to OGTT test in the protocol.

Correctness of actions in treatment phase Verify that there is an adjustment of therapy in case of blood glucose levels below 3.5 mmol/l (fasting and postprandial)–they are hypoglycemic, or fasting blood glucose values above 8 mmol/l or postprandial glucose levels above 10 mmol/l–both hyperglycemic.

Role Action correctness, related to reference values used for glucose monitoring.

Considered? No.

Why? First part is trivially (without formal verification) violated in the original protocol: see table with target values for blood glucose monitoring (postprandial values below 9 mmol/l are considered good).

Correctness of actions in insulin treatment Verify that in the event of insufficient correction (postprandial values stay above 10 mmol/l and preprandial values stay inside target area) a different insulin regimen is considered. This should be discussed with an internal medicine specialist.

Role Action correctness.

Considered? No.

Why? Not applicable. There is no mention in the protocol to simultaneous use of different types of glucose tests–postprandial and preprandial.

Safety of diet treatment Verify that oral medication is used when diet measures and increased physical activity have not been sufficient to reach the target blood glucose levels after 3 months.

Role Maximal duration of therapy.

Considered? No.

Why? Trivially (without formal verification) satisfied in the original protocol: see step 1 of treatment phase.

Safety of a combination of treatments Verify that a combination of all three classes of oral medication (sulphonylurea derivative, metformin and acarbose) is not prescribed.

Role Safety consideration.

Considered? No.

Why? Trivially satisfied in the original protocol: see steps 1 and 2 of treatment phase.

Different safety conditions Verify that non desired situations like diabetic foot ulcer, retinopathy, nephropathy, hypo- and hyperglycemia, are avoided.

Role Safety considerations.

Considered? No.

Why? Trivially satisfied in the original protocol: see e.g. the policies for concurrent diseases and hypoglycemic coma.

Correctness of management of patients with diabetic foot ulcer Verify that patients with diabetic foot ulcer that does not disappear within 2 weeks are referred to an adequate specialist.

Role Action correctness.

Considered? No.

Why? Trivially satisfied in the original protocol: see 'Consultation and referring' part.

Correctness of management of patients with retinopathy Verify that the lowering of blood glucose levels is more gradual in these patients.

Role Action correctness.

Considered? No.

Why? Not applicable. There is no mention to these particular patients in the protocol.

Correctness of management of patients with hypoglycemia using long-acting sulphonylureas Verify that these patients are monitored frequently.

Role Action correctness.

Considered? No.

Why? Not applicable. There is no mention to these patients in the protocol, nor to long-acting sulphonylureas.

Correctness of management of patients with hyperglycemia Verify that patients with hyperglycemia accompanied by drowsiness, dehydration or vomiting are addressed to a secondary care specialist and, in the event of hyperglycemic coma, referred to a hospital.

Role Action correctness.

Considered? No.

Why? Trivially satisfied in the original protocol: see 'Consultation and referring' part.

Correctness of management of patients with serum creatinine over 200 μmol/l or clearance below 30 ml/min Verify that these patients are referred to an internal medicine specialist or to a nephrologist.

Role Action correctness.

Considered? No.

Why? Trivially satisfied in the original protocol: see 'Consultation and referring' part.

Safety of management of some patients Verify that patients with kidney and liver insufficiencies, hypoxia with heart and vascular diseases, and heavy use of alcohol, do not get metformin.

Role Action correctness.

Considered? No.

Why? Not applicable. There is no mention to these patients in the protocol.

Correctness of management of patients under insulin treatment First, verify that patients using insulin with conditions of fever, vomiting or diarrhea take plenty of fluid and temporally increase therapy to prevent dehydration. Secondly, check that insulin is not discontinued nor reduced. Finally, these patients should undergo a careful control of blood glucose levels and fluid balance.

Role Action correctness.

Considered? No.

Why? Trivially satisfied in the original protocol (except last sentence): see 'Policy for concurrent diseases'.

Correctness of management of patients with overweight Verify that these patients get metformin in first place, rather that sulphonylurea derivatives or insulin.

Role Action correctness.

Considered? No.

Why? Trivially satisfied in the original protocol: see step 2 of treatment phase.

Correctness of management of patients with cardiovascular disease Ensure that these patients have a target blood pressure of 150/85 mmHg or lower.

Role Action correctness.

Considered? No.

Why? Trivially satisfied in the original protocol: see 'Treatment of other risk-factors of cardiovascular disease' part.

Correctness of sequencing of diagnosis and treatment actions Verify that the necessary data (e.g. weight, ketone bodies in urine etc) are collected after a diagnosis result of diabetes type 2, and that the right therapy is chosen after that.

Role Action correctness.

Considered? No.

Why? Trivially satisfied in the original protocol.

Correctness of sequencing of insulin treatment actions–1 In insulin therapy (in combination with oral antidiabetics or alone), a four-point day blood glucose curve must be performed before starting with insulin use and every day until the optimal insulin regime is established.

Role Action correctness.

Considered? No.

Why? Trivially violated in the Asbru protocol: insulin dose increase is performed every 7-10 days, according to our expert (this information was lacking in the original protocol) and four-point tests are only performed prior to these points.

Correctness of sequencing of insulin treatment actions–2 In insulin therapy, the starting doses must be 12 units before breakfast and 8 units of before the evening meal. This dosage must be modified every 2-3 days by 2-4 units based on the fasting and postprandial blood glucose values, until the target values have been reached.

Role Action correctness.

Considered? No.

Why? Trivially satisfied in the original protocol.

4.2 Incorporating background knowledge

As was discussed above, the definition of verification properties to consider in the case of diabetes type 2 guideline turned out to be a difficult issue, and we ended up with many properties of rather simple verification. Nonetheless, it is well known within the medical community that diabetes type 2 is actually a very complicated disease. This is caused by the fact that various metabolic control mechanisms are involved as well as by the many different organ systems, such as the cardiovascular tract and the renal system, which may be affected by diabetes. Here we simply focus on the derangement of glucose metabolism in diabetic patients. The reader should be aware, however, that this is only one out of a plethora of different derangements which are associated with this condition. In order to assist general practitioners in the management of this complicated disease in patients, being able to consult a guideline is really essential.

The two observations mentioned above seem contradictory: many of the properties we considered seem to be simple and easy to verify, yet the guideline deals with a disorder that is very complicated. Our first conclusion was, therefore, that this guideline is really well designed so that it is not possible to find major flaws. This contrasts with the jaundice guideline, which seems to be of much lower quality. Our second conclusion was that even the well-designed diabetes type 2 guideline might still contain non-trivial flaws. However, in order to be able to find such flaws a medically speaking much deeper analysis would be required.

The results of such a deeper analysis are discussed in the following sections. It must be noted that this work is preliminary in nature, and is included here only as one of the possible inputs for future work in the area of formal verification of medical guidelines and protocols.

4.2.1 Diabetes type 2 background knowledge

Figure 1 summarises the most important mechanisms and drugs involved in controlling the blood levels of glucose. The protein hormone insulin, which is produced by the *B cells* in the Langerhans islets of the *pancreas*, has the following major effects:

- it increases the uptake of glucose by the liver, where it is stored as glycogen, and inhibits the releases of glucose from the liver;
- it increases the uptake of glucose by insulin-dependent tissues, such as muscle and adipose tissue.

At some stage in the natural history of the disorder diabetes mellitus type 2, the level of glucose is too high (hyperglycaemia) due decreased production of insulin by the B cells. A popular hypothesis explaining this phenomenon is that target cells have become insulin resistant, which with a delay causes the production of insulin by the B cells to raise. After some time, the B cells become exhausted, and they are no longer capable of meeting the demands for insulin. As a consequence, hyperglycaemia develops.

Treatment of diabetes type 2 consists of:

- Prescription of *sulfonylurea* (SU) drugs, such as tolbutamid. These drugs stimulate the B cells in producing more insulin, and if the cells are not completely exhausted, the hyperglycaemia can thus be reversed.
- Prescription of *biguanides* (BG), such as metformin. These drugs inhibit the release of glucose by the liver.
- Prescription of *α-glucosidase inhibitors*. These drugs inhibit (or delay) the absorption of glucose by the intestines.
- Injection of *insulin*. This is the ultimate, causal treatment.

As insulin can, in contrast to the other drugs which are normally taken orally, only be administered by injection, general practitioners prefer to wait with prescribing insulin as long as possible. Thus, the treatment part of the diabetes type 2 guideline mentions that one should start with prescribing oral antidiabetics (SU or BG). Two of these can also be combined if taking only one has

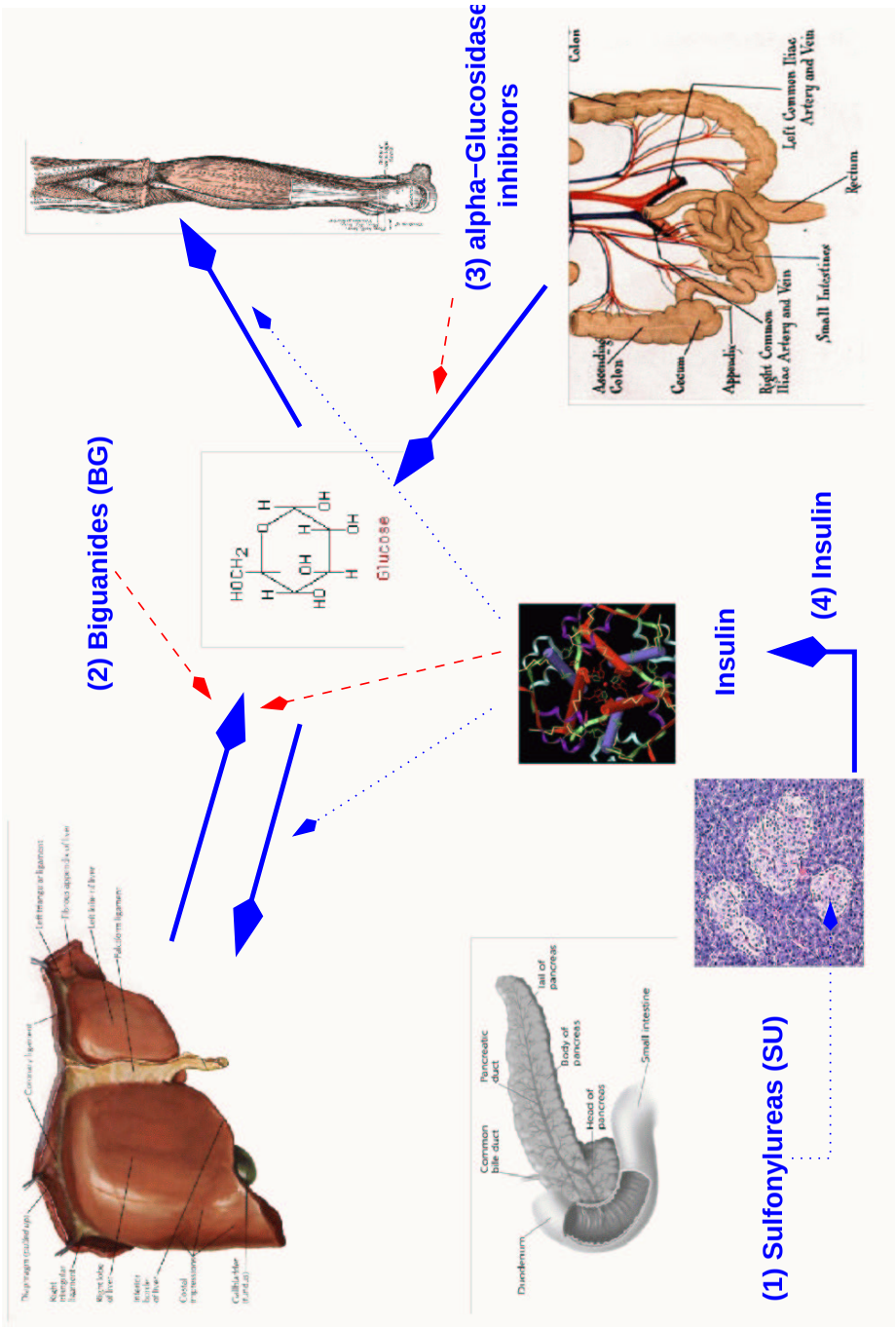


Figure 1: Summary of drugs and mechanisms involved in the control of blood levels of glucose.

insufficient glucose-level decreasing effects. If treatment is still unsatisfactory, the guideline suggests either (1) to add insulin, or (2) to stop with the oral diabetics entirely and to start with insulin.

From a medical point of view, advice (1) above is somewhat curious. Since the oral antidiabetics are not effective anymore, the conclusion must be that the B cells must be exhausted. Under these circumstances, it does not make a lot of sense to prescribe an SU drug. Prescription of a BG (or α-glucosidate inhibitor) is still justified, as by adding an oral antidiabetic to insulin, the number of necessary injections can be reduced from twice a day to once a day.

It may be argued that in this way the guideline violates one of the principle of good practice in medicine, which is to prescribe as few drugs as possible, taking into account costs and side-effects of drugs. Obviously, finding a potential flaw as the one discussed above requires a lot of medical background knowledge; simply reading through the text without a deep biomedical understanding would not allow anyone to spot problems such as these.

4.2.2 Formalisation of good practice medicine

All the above make one wonder whether the knowledge deployed above in spotting flaws in a well-designed medical guideline can be formalised. If the answer to this question is positive, then such knowledge could be used as a basis for work in formal verification of medical guidelines.

As is clear from the description above, two types of knowledge were involved in detecting the violation of good medical practice:

- Knowledge concerning the (patho)physiological mechanisms underlying the disease (here diabetes type 2). In the case of diabetes type 2 the knowledge involved is causal in nature.

- Knowledge concerning good practice in treatment selection.

Below we present some preliminary ideas on how such knowledge may be formalised so that it can be used by a theorem prover such as KIV. Again, note that the ideas presented here have only been included as a possible direction for future work.

We are interested in the prescription of drugs, taking into account their mode of action. This can be formalised in logic, although somewhat naively, as follows:

$$d \rightarrow (m_1 \wedge \cdots \wedge m_n)$$

where d is the name of a drug or possibly of a group of drugs indicated by a predicate symbol (e.g. SU(x), where x is universally quantified; and SU(tolbutamid) for individual drugs), and m_k is a mode of action, such as decrease of release of glucose from the liver.

The modes of actions m_k can be combined, together with *intentions* n (achieving normoglycaemia for example), and particular patient *conditions* c and *requirements* r for the mode of actions to be effective:

$$(m_{i_1} \wedge \cdots \wedge m_{i_m} \wedge c \wedge r) \rightarrow n$$

Good practice medicine can then be formalised as follows. Let $\mathcal{T} \subseteq \{d_1 \ldots d_p\}$ be a *treatment*, C a collection of patient conditions, R a

⟨*plan*⟩(. . .)
begin
  await ⟨*plan*⟩-filter(. . .);
  ⟨*plan*⟩-state := activated;
  pbreak
    ⟨*plan-body*⟩
  if ⟨*plan*⟩-abort(. . .);
  if ⟨*plan*⟩-abort(. . .)
    then ⟨*plan*⟩-state := aborted
    else ⟨*plan*⟩-state := completed
end;

axioms

filter-def: ⟨*plan*⟩-filter(. . .) ↔ . . . ;
abort-def: ⟨*plan*⟩-abort(. . .) ↔ . . . ;

Figure 2: Plan Observation in KIV

collection of requirements, and N a collection of intentions for the treatment. A treatment $\mathcal{T}$ is in accordance to good-practice medicine if:

- $\mathcal{T} \cup C \cup R \models N$, and
- $\mathcal{T}$ is subset minimal, with the least preferred drugs omitted if possible.

In the context of abductive reasoning, subset minimality is often used in order to distinguish between various solutions; a criterion such as this is referred to in literature as *Occam's razor*.

With this formalisation in mind, and using the current formalisation and verification approach in Protocure, it is possible to formally establish that there is a flaw in the guideline as under certain conditions, three drugs are prescribed whereas only two should have been prescribed according to the good practice principle. However, the fact that only two drugs should have been prescribed is a specific derived property of this guideline. It would have been much more interesting, but also more difficult, to input the general principle into a theorem prover such as KIV and find that this *general* principle is violated. As is argued above, it is very likely that a substantial amount of medical background knowledge, such as concerning diabetes type 2, will be required to achieve this goal. We think that collecting such knowledge might be one of the challenging directions for future research.

5 Spin-Offs

5.1 Formalisation improvements

5.1.1 Patterns for Asbru plans

Our verification effort has helped to optimize the translation of Asbru plans. Our goal has been to improve readability of plans in KIV and to better distinguish between the actual plan and the necessary overhead to model Asbru plan

behaviour in general. In Figure 2 a translation pattern for simple plans without complete and continuation condition is given. With the definition of predicates $\langle plan \rangle$-filter(...) and $\langle plan \rangle$-abort(...) the general Asbru plan pattern has been separated from the definition of conditions for concrete plans. For example, to translate plan Observation the following predicates are defined:

filter-def: ob-filter(pd) $\leftrightarrow$ get-bilirubin(pd .tsb) = observation;
abort-def : ob-abort(pd) $\leftrightarrow$ get-bilirubin(pd .tsb) $\neq$ observation;

The separation of general Asbru plan behaviour from concrete plan definition also helps to recognize patterns in proofs. Executing an Asbru plan always implies to execute the filter condition first, then to assign the $\langle plan \rangle$-state variable next and so on.

In the future the translation can be further improved. For example the definition of the plan body could also be separated from the general pattern. Furthermore, more complex plans with continuation conditions can also be considered. A pattern for translations also gives rise to more automatic translation from Asbru to KIV and helps to close the gap between both.

5.1.2 Patterns for Intentions

In our case study of Jaundice we have verified a number of intentions. For their medical relevance, we are interested in the results of these specific intentions. That means whether the intention holds, or under which particular circumstances the intention holds. We see verification of intentions as a generally useful approach to improve medical protocols, therefore we are interested in general guidelines for formalising intentions from the Asbru language into temporal formulae. From this viewpoint, we first have to identify different types of intentions, then we have to identify schemas of temporal formulae for these types of intentions. Those schemas have been used in our case study of Jaundice. During the verification process it turned out that we needed to improve the general schemas, as we will discuss in this section.

Types of intentions In Asbru we can distinguish several types of intentions:

- achieve, maintain or avoid
- intermediate or final
- state or action
- intention conditions with or without time annotations
- intention of a child plan or of a parent plan

These types are orthogonal. An example of an intention schema is: achieve intermediate state $condition_i$ with $time\text{-}annotation_j$ from a child plan (an individual plan).

In this one-year Protocure project, a choice had to be made on the most interesting cases. Therefore, we took only into account those types of intentions which were the most medically relevant ones in the Jaundice case study (see Section 3). Our focus is on the following intentions schemas:

- maintain intermediate state (child plan, without time-annotation)
- maintain intermediate state (child plan, with time-annotation)
- achieve final state (child plan, without time-annotation)
- avoid intermediate state (child plan, without time-annotation)
- hierarchy (intention of a parent, without time-annotation)

Below we describe their counterpart as a temporal formula schema.

maintain intermediate state *condition*$_i$

$$\vdash \text{plan}\#(\ldots) \rightarrow \Box \; \textit{plan-state} \text{ .activated} \rightarrow \textit{condition}_i$$

For every step in the program (plan#(...) →), the required condition (*condition*$_i$) always holds from now on (□) under the condition that the plan is in activated state (*plan-state* .activated →).

maintain intermediate state *condition*$_i$ *time-annotation*$_j$ We can use the same schema as above for this intention type. Only the complexity of the condition will increase because of the time annotation.

achieve final state *condition*$_i$

$$\vdash \text{plan}\#(\ldots) \rightarrow \Box \; \textbf{laststep} \wedge \textit{plan-state} \text{ .completed} \rightarrow \textit{condition}_i$$

For every step in the program (plan#(...) →), it always holds (□) that, under the condition that it is the last step (**laststep**) and the plan is in completed state (*plan-state* .completed), the required condition (*condition*$_i$) holds.

avoid intermediate state *condition*$_i$

$$\vdash \text{plan}\#(\ldots) \rightarrow \Box \; \textit{plan-state} \text{ .activated} \rightarrow \neg \; \textit{condition}_i$$

For every step in the program always, if the plan is activated, the condition *condition*$_i$ does not hold.

hierarchy The distinction between an intention of a child (individual) plan and an intention of a parent plan has nothing to do with the formulation of the intention itself. It only plays a role in the way of proving the property. In case of a parent plan, we can use the proofs of the intentions of its children. One of KIV's features is the possibility of introducing lemmas. This means that the proof of a parent intention can use the lemmas of the intentions of its children.

Formulation improvements of intentions During the proof attempts of an intention of the type "maintain intermediate state *condition*$_i$" (child, with time-annotation), several improvements of the formalisation had to be made. This resulted in an improvement of the general schema. First we describe these improvements. We continue with an illustration of these improvements for a concrete intention from the Jaundice protocol that has been proven.

The general improvements are:

1. Be aware of environment assumptions. That means there is a need for specifying which variables can be changed by the environment, and which can not.

2. In the general schema we require the condition for all steps. However in the last step it is not possible to refer to primed variables. Therefore we need a case distinction: we are in the last step or the $condition_i$ holds.

3. The $condition_i$ holds when the plan is activated until the plan is not activated anymore.

The above statements result in the following schema:

$$
\begin{array}{lll}
 & \text{plan\#}(;\ldots) & \\
\rightarrow & & \textit{environment-assumption} \\
 & \textbf{unless laststep} & \\
 & \rightarrow (\Box & \textbf{laststep} \\
 & & \vee \quad \textit{pti-state} \neq \text{activated} \wedge \textit{pti-state}' = \text{activated} \\
 & & \rightarrow \qquad \textit{time-annotation} \\
 & & \qquad \rightarrow \textit{condition}_i \\
 & & \textbf{unless } \textit{plan-state}' \neq \text{activated}
\end{array}
$$

For every step in the program (plan#(;...)), under the assumptions in the environment-assumption part, it always holds from now on that either we are in the last step (**laststep**) or, under the condition that the plan is activated and the time-annotation has not elapsed, then the required $condition_i$ holds.

The following intention of the Phototherapy-intensive plan serves to illustrate this:

$$
\begin{array}{l}
\text{MAINTAIN INTERMEDIATE STATE} \\
\quad (\textit{TSB-decrease} = \text{yes})\ ([4h, -][-, 6h][-, -], \textit{self}) \\
\text{AND}\ (\textit{TSB-change} \geq 10)\ ([4h, -][-, 6h][-, -], \textit{self}).
\end{array}
$$

The components of the formula are:

- environment assumption:

$$
\begin{array}{ll}
 & \textit{pti-state}'' = \textit{pti-state}' \\
\wedge & \textit{pti-acttime}'' = \textit{pti-acttime}' \\
\wedge & \textit{time}'' = \textit{time}' + 1 \\
 & \textbf{unless laststep}
\end{array}
$$

  This means that only the time will be changed by the environment, and *pti-state* and *pti-acttime* will not.

- time annotation:

$$4 \leq \textit{time} - \textit{pti-acttime}\ .\text{value} \wedge \textit{time} - \textit{acttime}\ .\text{value} \leq 6$$

  The variable *pti-acttime* is the starting time of the plan phototherapy-intensive. The first part implies starting after 4 hours of the activation time of phototherapy-intensive plan. The second part implies finishing before 6 hours after the activation time of phototherapy-intensive plan.

- condition of the intention:

$$\begin{aligned} &\text{get-decrease}(pd\ .\text{tsb}, pti\text{-}acttime\ .\text{value}) \\ \wedge\ &\text{get-change}(pd\ .\text{tsb}, pti\text{-}acttime\ .\text{value}) \geq 10 \end{aligned}$$

These components together with the general schema of this intention type result in the formula:

$$\begin{aligned} &\text{phototherapy-intensive}\#(; ptnp\text{-}acttime, pd, time, pti\text{-}acttime, pti\text{-}state) \\ \rightarrow\ &pti\text{-}state'' = pti\text{-}state' \\ &\wedge\ pti\text{-}acttime'' = pti\text{-}acttime' \\ &\wedge\ time'' = time' + 1 \\ &\textbf{unless laststep} \\ \rightarrow\ &(\Box \quad \textbf{laststep} \\ &\vee \quad pti\text{-}state \neq \text{activated} \wedge pti\text{-}state' = \text{activated} \\ &\rightarrow \quad 4 \leq time - pti\text{-}acttime\ .\text{value} \wedge time - acttime\ .\text{value} \leq 6 \\ &\rightarrow \quad \text{get-decrease}(pd\ .\text{tsb}, pti\text{-}acttime\ .\text{value}) \\ &\wedge\ \text{get-change}(pd\ .\text{tsb}, pti\text{-}acttime\ .\text{value}) \geq 10 \\ &\textbf{unless}\ plan\text{-}state' \neq \text{activated}) \end{aligned}$$

The attempt to formally prove intentions in KIV has forced us to be more precise, and to refine the original schema for intentions. We expect similar refinements are required for the other intention types.

5.2 KIV improvements

Elaborate tool support is essential for the efficient verification of properties. Within our one year assessment project, we were able to formally verify only a small number of properties. This is due to the fact that tool support for verifying concurrent systems and temporal logic had to be developed in parallel to the Protocure project.

Initially, a basic calculus for standard parallel programs and temporal logic was available in KIV, which was enough to do simple proofs for simple programs. Three major items had to be improved during the project:

- Support for more complex operators

  In order to keep the gap between Asbru plans and their KIV representation as small as possible, a number of complex program constructs have been implemented (see [2]). These constructs required additional proof rules.

- High-level proof rules

  The rules of the basic calculus were sufficient to manually prove simple properties for simple plans. For example, these rules have been used to prove the termination of Phototherapy-intensive and other simple plans in the Jaundice protocol.

  More complex proofs would have involved too many basic proof steps. This is why higher level proof rules were required. Instead of rewriting each program construct separately, higher level rules rewrite a whole group of constructs within a single step. Doing the proofs for the Jaundice protocol was very important to develop, implement and fine tune these higher level rules.

- Automation

  More complex properties and plans require more automatic proofs. Proving the indicator for the Regular-treatments plan in the Jaundice protocol (see Appendix B) involves thousands of proof steps. Also, correcting detected errors invalidates proofs. More automatic proofs can be repeated more easily.

  In principle, the proof strategy of symbolic execution can be automated to a very large extent. However, the design and implementation of heuristics which automatically apply rules is not trivial. The proofs for the Jaundice protocol gave very important feedback to the developers of the KIV system on how to make the heuristics more powerful and user friendly. Work on this issue is still not finished.

Except for the first item, these improvements are of a general nature. High-level rules and heuristics were implemented such that they are applicable to parallel programs and temporal logic in general.

6 Future Challenges

6.1 Improving translation patterns

The patterns to translate Asbru plans to KIV formulas as described in Section 5.1.1 should be further improved. The availability of better patterns for translation would give rise to a more automatic and less error-prone translation from Asbru to KIV, and would thus narrow the gap between both languages even further. This is seen by the project team as one of the necessary steps in order to make verification of medical guidelines and protocols practically feasible.

While verifying Asbru plans using KIV, certain sequences of proof steps did reappear again and again. This is not surprising given the similarity in structure of Asbru plans and subplans, and also of similarity in content of plans. Clearly, if these sequences of proof steps could be generalised such that they could be used as a sort of proof tree schemata, to be employed for proving properties of different plans, possibly even regarding different protocols, then this would give rise to huge savings in terms of verification efforts. Some ideas in this direction have been developed during the project, but need to be further elaborated to become practically useful.

6.2 Improving tool support

Currently the KIV tool is not powerful enough to do verification of large concurrent systems. It turned out that formulating lemmas in constructing a modular proof is difficult. Quite often during the main proof the lemmas abstracting the behaviour of subplans need to be adapted, resulting in a repetition of proofs. During our case study, the following two additional ingredients to the proof strategy of KIV were dearly missed:

- Reuse of proof attempts

  After errors have been corrected, the KIV system currently allows for the automatic reuse of proof attempts for sequential programs and predicate

logic. This feature also needs to be implemented for parallel programs and temporal logic.

- Generalisation

  It should be possible to automatically analyse completed proof branches and to extract those parts of the formulae which were really needed to prove the sub goal. This gives rise to lemmas which can be applied to similar situations.

  This feature is very important, because the current proof strategy for verifying concurrent systems results in many similar cases.

6.3 Adding medical background knowledge

While not part of the verification effort *per se*, during verification it became clear that some of the medically speaking more interesting flaws in a guideline can only be detected if biomedical background knowledge is added to the formalised guideline. This was illustrated in Section 4.2 for diabetes mellitus type 2. It is at the moment not completely clear, not even for diabetes mellitus type 2, which types of knowledge, for example causal or structural knowledge, are required for this purpose. As was already seen in the case of diabetes type 2, it seems likely that this knowledge resides at two levels: at the object level, which concerns knowledge of the structure and the functional mechanisms involved in the disease, and at the problem-solving level, involving knowledge about good medical practice and disease management. Even though it may not be easy to elucidate, formalise and formally check these types of knowledge, such an endeavour could have an immediate impact in the medical field, and would therefore be an interesting future line of research.

7 Discussion

7.1 Benefits of formal methods

One possibility to integrate formal methods into the development of guidelines is as follows:

Medical Expert: Here is a new guideline. We have developed the guideline with an intention in mind. If our guideline is used, is the intention achieved?

Verifier: No, here is a counter example describing a case, where the intention is not achieved.

Medical Expert: Ah, right. I forgot to mention that in this case, the intention need not be achieved. I will mention this explicitly within our guideline.

Verifier: OK, then I will also assume that for the purpose of verifying the intention, the problematic case will never occur.

However, for your intention to hold we need to assume even more. If I make the following assumptions, the intention is provable.

Medical Expert: Most of the assumptions are reasonable, because that is what is true in most cases. The assumptions are violated only in exceptional cases and here the intention may be violated also.

But, some of the assumptions are too strong. The intention should be achieved even if the assumption does not hold.

Verifier: Then you need to improve your guideline such that it incorporates the crucial cases.

Medical Expert: OK, here is an improved protocol. Is the intention achieved?

Verifier: Yes, I have systematically analysed all possible cases. Now the intention holds under the remaining assumptions.

Medical Expert: OK, then our guideline is of good quality. Also, the remaining assumptions somehow explicitly define exceptional cases. Probably the assumptions help in identifying patients where the treatment does not run normally.

7.2 Advantages of formal methods

Medical Expert: Why did we spend so much effort in this formal verification stuff? That there are cases in which the intention is not achieved can be seen just by taking a closer look at the informal guideline.

Verifier: True, this is also part of the advantage of formal methods. If you try to formally model and verify a guideline you are forced to progress slowly and think of details. That is why already while writing the *formal model*, a lot of flaws in the original guideline are detected.

Finally, *formal verification* uncovers problematic cases in a systematic fashion. Other methods normally only reveal the more obvious cases. With formal methods you can be sure that all of the cases have been analysed.

7.3 Formal verification harder than expected

Medical Expert: Well. But why did you only formally verify the Jaundice guideline and even this guideline you only considered in part? This is such a small bit of a guideline. We need to develop guidelines much faster.

Verifier: Without good tool support, the use of formal methods is very difficult. Support for the verification of guidelines in KIV currently is very rudimentary. We have underestimated the effort of building sophisticated tool support for the verification. But already during this one year, we have significantly increased our efficiency in verifying guidelines. Also, we came up with lots of ideas to improve our tools even further, but found no time to implement them yet.

```
phototherapy-intensive# (ptnp-acttime, pd, time, pti-acttime, pti-state)
begin
   await pti-filter(pd);
   pti-acttime := mk-value(time);
   pti-state := activated;
   pbreak
      prescribe-intensive-phototherapy#(; pd, time)
   if pti-abort(pd, time, pti-acttime);
   if pti-abort(pd, time, pti-acttime) then
      ⟨pti-state := aborted;
       pd := set-under-phototherapy(pd, false)⟩
   else
      pti-state := completed
end;

prescribe-intensive-phototherapy# (pd, time)
begin
   pd := set-under-phototherapy(pd, true);
   var time0 = time, dt = ? in await time0 + dt ≤ time;
   pd := set-under-phototherapy(pd, false)
end
```

Figure 3: Definition of Phototherapy-intensive plan

A Example Proof

As an example proof we will present the verification of an intention for the Phototherapy-intensive plan of the Jaundice protocol, which is given in Figure 3. This is a rather simple proof, because the intention is implemented as an abort condition. A more complicated proof is described in Appendix B.

For verifying the intention, the proof obligation of Figure 4 needs to be satisfied. The tree of the completed proof is depicted in Figure 5. The proof strategy is symbolic execution with induction. Basically, we need to execute Phototherapy-intensive to prove the property. There is a strong resemblance between the program structure (see Figure 3) and the proof. This resemblance can be discovered by looking at the most interesting proof steps (numbers below refer to nodes in Figure 5):

3) First step of Phototherapy-intensive is executed. In premise 5, filter condition is satisfied. In premise 43, it does not hold at first.

6) Second step is executed and *pti-state* is set to activated. Now that the plan has been activated, the intention needs to be satisfied. This is proven in branch 7. Branch 25 is artificial. Improving the heuristics should eliminate this branch.

8) Third body of plan Phototherapy-intensive is executed. Executing the body requires calling Prescribe-intensive-phototherapy (premise 9). If Phototherapy-intensive is immediately aborted, the intention trivially holds (premise 10). In premise 11, the abort condition does not trigger and we continue execution.

$$
\begin{array}{l}
\quad \textit{/* Phototherapy-intensive plan */} \\
\quad \text{phototherapy-intensive\#}(\ldots) \\
\wedge\ \textit{/* environment assumption */} \\
\qquad \textit{pti-state}'' = \textit{pti-state}' \\
\qquad \wedge\ \textit{pti-acttime}'' = \textit{pti-acttime}' \\
\qquad \wedge\ \textit{time}'' = \textit{time}' + 1 \\
\quad \textbf{unless laststep} \\
\rightarrow\ \textit{/* intention */} \\
\quad \Box \quad \textbf{laststep} \\
\qquad \vee \quad \textit{pti-state} \neq \text{activated} \wedge \textit{pti-state}' = \text{activated} \\
\qquad\quad \rightarrow \quad \textit{/* time annotation */} \\
\qquad\qquad\qquad 4 \leq \textit{time} - \textit{pti-acttime}\ .\text{value} \\
\qquad\qquad\qquad \wedge\ \textit{time} - \textit{pti-acttime}\ .\text{value} \leq 6 \\
\qquad\qquad\quad \rightarrow\ \textit{/* property */} \\
\qquad\qquad\qquad\quad \text{get-decrease}(\textit{pd}\ .\text{tsb}, \textit{pti-acttime}\ .\text{value}) \\
\qquad\qquad\qquad\quad \wedge\ \text{get-change}(\textit{pd}\ .\text{tsb}, \textit{pti-acttime}\ .\text{value}) \geq 10 \\
\qquad\qquad\ \textbf{unless}\ \textit{pti-state}' \neq \text{activated}
\end{array}
$$

Figure 4: Proof obligation for intention of Phototherapy-intensive

13) Executing Prescribe-intensive-phototherapy can take an arbitrarily long time. Therefore induction is required.

15) One step of Prescribe-intensive-phototherapy is executed. Again, the property trivially holds, if the parent plan Phototherapy-intensive aborts (premises 16, 21). In premise 17, Prescribe-intensive-phototherapy completes; in premise 22, it does not.

20) One final step executes the remaining conditional and the assignment of variable *pti-state*, and finishes the proof.

24) If Prescribe-intensive-phototherapy does not terminate, we can apply the induction hypothesis to finish the proof.

43) The filter condition of Phototherapy-intensive is not satisfied. Execution could continue waiting forever for the condition to be satisfied. Therefore induction is required here.

47) In this step, we again try to execute the await statement corresponding to the filter condition. In branch 48, the condition finally holds. In branch 87, it does not. Here we apply induction (node 89) to finish the proof.

48) This branch is basically a repition of the proof for node 5. Improved heuristics may eliminate this branch.

Strong resemblance between proof and program structure gives very intuitive proofs. With improved heuristics we have been able to construct a proof with roughly 1/3 of the proof steps.

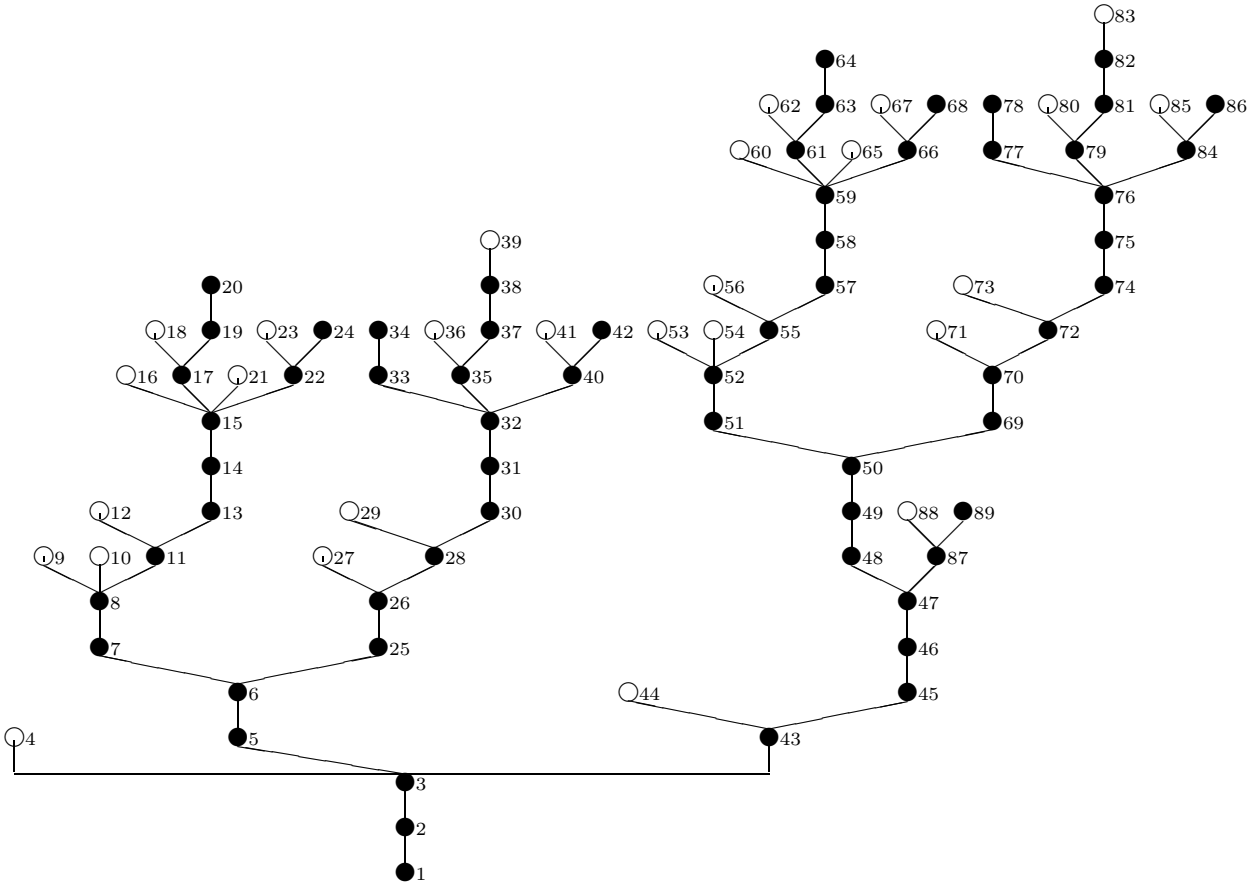


Figure 5: Proof tree for intention of Phototherapy-intensive

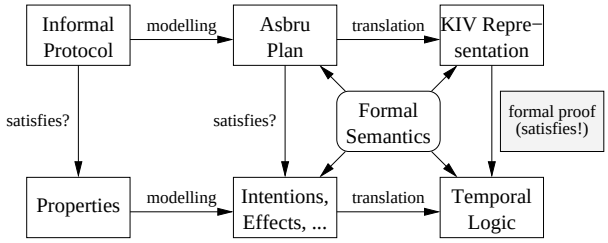
B Report on Verification of Indicator

B.1 Introduction

In this document, we report on our experiences with formally verifying a medical guideline (or protocol). This work is part of an IST European project called Protocure which has the aim of assessing if the quality of medical guidelines can benefit from formal methods. Our basic assumption is that guidelines can be seen as computer programs. Therefore, development methods which ensure high quality of computer programs can also be applied to medical guidelines. With formal methods, very high quality computer programs can be developed. Can the application of formal methods to medical guidelines also improve their quality?

To answer this question we have formalised a medical guideline from the American Association of Pediatrics, intended for the management of newborn jaundice, and verified one of the criteria (indicators) established for the management of this disease. In the verification tasks we made use of the interactive theorem prover KIV. Medical guidelines basically define a concurrent system. Proof support for concurrent systems has recently been added to the KIV verifier. Therefore another goal of this work has been to evaluate and improve the current support offered by KIV for verifying concurrent systems.

The idea of our approach can be depicted as follows:



We would like to know if the informal Jaundice guideline (see Sect. B.2.1) satisfies one of the MAJIC indicators (see Sect. B.2.2). In a first step, we have constructed a semi-formal model of the Jaundice guideline. The model is written in Asbru, a knowledge representation language especially designed for medical protocols (see Sect. B.3.1). In a second step, this model has been formalised into KIV, an interactive theorem prover. Asbru plans have been mapped onto parallel programs, which in this way can be used as input format to KIV (see Sect. B.3.2). The indicator has been then translated to a formal property in Temporal Logic (see Sect. B.3.3). The main part of this paper reports on our experiences with formally verifying the temporal logic property in KIV (see Sect. B.4).

B.2 Documents

B.2.1 Jaundice guideline

Jaundice (or hyperbilirubinemia) is a common disease in newborn babies. Under certain circumstances, elevated bilirubin levels may have detrimental neurological effects. In many cases jaundice disappears without treatment but sometimes needs phototherapy to reduce the levels of total serum bilirubin (TSB), which

indicates the presence and severity of jaundice. In a few cases, however, jaundice is a sign of a severe disease.

The jaundice protocol of the American Association of Pediatrics (AAP) [1] is intended for the management of the disease in healthy term[2] newborn babies. The main reason for choosing this protocol was that it is considered a high-quality protocol: the jaundice protocol of the AAP is included in the repository of the National Guideline Clearinghouse[3].

The guideline is a 10 pages document. It consists of an evaluation (or diagnosis) part and a treatment part, to be performed in sequence. During the application of the protocol, as soon as the possibility of a more serious disease is uncovered, the recommendation is to exit without any further action. The rationale behind this is that the protocol is exclusively intended for the management of jaundice in healthy newborns. An important part of the protocol is the table used to determine the adequate treatment from the TSB value and the age of the infant.

B.2.2 MAJIC indicators

The MAJIC (Making Advances against Jaundice in Infant Care) Steering Committee introduces in [3] a set of 11 indicators that define the criteria to be considered in the management of newborn jaundice. These indicators should hold for any guideline intended for the management of this disease, including the AAP one. The developed indicators are a little more lenient than the AAP guideline.

An example for an indicator follows:

INCLUSIONS If infant 25-48 hours old

CRITERIA

- Repeat bilirubin drawn within 24 hours if serum bilirubin level $\geq$ 14 mg/dl
- Phototherapy initiated within 24 hours if serum bilirubin level $\geq$ 16 mg/dl

For our case-study we studied the following indicator:

INCLUSIONS If any phototherapy initiated

CRITERIA No more than one serum bilirubin level drawn after phototherapy is discontinued

B.3 Formalisation

In this document, we sketch the Asbru model of the Jaundice guideline and of the above indicator. For more details on the former, see [4].

[2] Defined as 37 completed weeks of gestation.

[3] `http://www.guideline.gov`

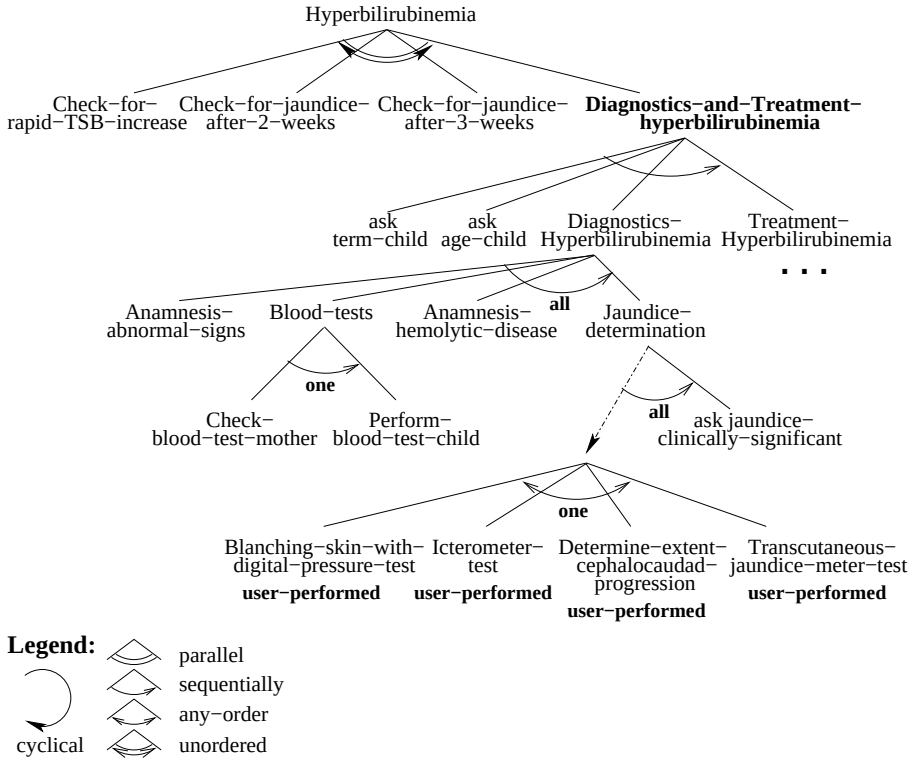


Figure 6: Overview of the jaundice model in Asbru, with details of the diagnostics part.

B.3.1 Guideline in Asbru

Like the AAP protocol, the Asbru version has as main components a diagnostics part and a treatment part. It is made up of about 40 plans and has a length of 16 pages in an intermediate Asbru notation. Figures 6 and 7 show together the overall structure of the jaundice protocol as a hierarchy of Asbru plans. In these figures, the links represent the decomposition of plans into steps or subplans, and the type of decomposition is given by different kinds of arrows. Besides, the waiting strategy is indicated either with keywords (e.g. **one** or **all**) or with bold font (e.g. the plan "Diagnostics-and-treatment-hyperbilirubinemia" is a compulsory part, which means it should be completed before the "Hyperbilirubinemia" can complete). Finally, notice that figures do not include all the plans in the protocol, for readability reasons.

The most important entry point of the protocol is the plan "Diagnostics-and-treatment-hyperbilirubinemia"–the three "Check-for-..." plans are Asbru artifacts to model, respectively, a continuous monitoring of TSB level and two check-ups at temporally specified intervals. The plan "Diagnostics-and-treatment-hyperbilirubinemia" is divided into a diagnostics subplan and a treatment subplan, to be executed sequentially. In the rest of the section we give an overview of these parts, and describe their main subplans. For more details, see [4] and/or the XML version of the protocol in `http://www.protocure.org/`.

The diagnostics stage (in figure 6) is in turn subdivided into four sequential

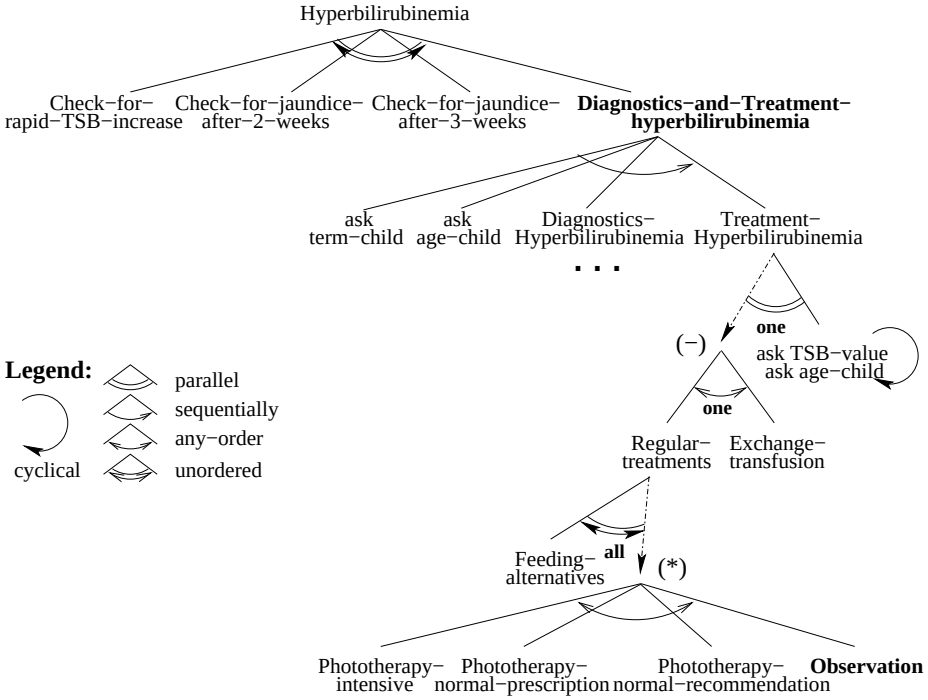


Figure 7: Overview of the jaundice model in Asbru, with details of the treatment part.

plans. It performs actions looking for a possible pathologic reason and investigating whether jaundice is significant, in this order. In principle it is considered that no pathologic reason exists. The plan only completes when all the actions have been completed, hence the keyword **all** appears as waiting strategy. The actions consist of a series of anamnesis questions and blood tests, to determine the possibility of other (more serious) diseases.

The treatment phase consists of two parallel parts, namely the actual treatments and a cyclical plan asking for the input of new age and TSB values every 12 to 24 hours. Regarding the treatments (label (-) in figure 7), either the regular ones ("Regular-treatments") or an exchange transfusion ("Exchange-transfusion") can take place depending on the bilirubin level. Besides, although it is not shown in figure 7, an exchange transfusion should be applied in case "Regular-treatments" aborts, which would indicate a failure of these regular treatments.

The "Regular-treatments" plan contains the main treatment procedure. It consists of two parts to be performed in any possible order (unordered): the study of feeding alternatives and the different therapies (see label (*) in figure 7). The plans in group (*) can be tried in any order, one at a time. The intentions of "Regular-treatments" plan are both avoiding toxic bilirubin levels and attaining normal (or observation) ones at the end. The plan completes when the feeding alternatives and the therapies complete (waiting strategy **all**). The latter in turn depends on the completion of observation (compulsory plan in bold). It aborts when either bilirubin raises to transfusion levels or intensive phototherapy fails

```
regular-treatments# (...)
var pti-state = inactive, ptnp-acttime = ⊥, ptnp-state = inactive in
begin
   /* filter condition */
   await rt-filter(pd);
   rt-state := activated;
   pbreak
      /* plan body */
         feeding-alternatives#(...)
      ||_s var ob-state = inactive in
         pbreak
            while true do
                  phototherapy-intensive#(...)
               or phototherapy-normal-prescription#(...)
               or phototherapy-normal-recommendation#(...)
               or observation#(...)
         if rt-continuation(ob-state)
   if rt-abort(pd, time, pti-acttime);
   if rt-abort(pd, time, pti-acttime)
   then rt-state := aborted
   else rt-state := completed
end
```

Figure 8: Plan Regular-treatments represented in KIV

to reduce them sufficiently, pointing to a pathologic reason.

B.3.2 Guideline in KIV

In order to verify the formal model in the theorem prover KIV, the model needs to be translated to a format known to KIV. For this purpose, parallel programs have been chosen, because Asbru plans can be seen in principle as a concurrent system. Each Asbru plan is translated into one procedure. As an example, Figure 8 contains the resulting procedure for the Regular-treatments plan. The procedure starts by evaluating the filter condition rt-filter(...). If the filter condition is satisfied, the body of the plan is executed. Execution of the body is guarded by the abort and continuation conditions rt-abort(...) and rt-continuation(...). In the body, Feeding-alternatives is executed in parallel (... $||_s$ ...) to a choice of treatments (... **or** ...). Additional variables *rt-state*, *pti-state*, etc. are introduced to capture the Asbru execution state.

B.3.3 Indicator in Temporal Logic

The indicator is translated to a temporal formula (see Figure 9). Always, if phototherapy is discontinued (the patient is currently under phototherapy but, in the next step, he/she is not under phototherapy anymore), we remember the current history of TSB readings TSB_0. If, in the future, another measurement is taken (the history TSB_1 is different from the original history), then there will be no more TSB measurements (the TSB history *pd* .tsb will always be the same). Thus only one TSB measurement is allowed, if phototherapy is discontinued.

$$
\begin{array}{l}
/^{*}\ \textit{indicator}\ ^{*}/ \\
\Box \quad \textbf{laststep} \\
\quad \vee \qquad \textit{pd}\ .\text{under-phototherapy} \wedge \neg\ \textit{pd}'\ .\text{under-phototherapy} \\
\qquad\quad \wedge\ \textit{pd}\ .\text{tsb} = \textit{TSB}_0 \\
\qquad \rightarrow \Box\ \textit{pd}\ .\text{tsb} = \textit{TSB}_1 \wedge \textit{TSB}_1 \neq \textit{TSB}_0 \rightarrow \Box\ \textit{pd}\ .\text{tsb} = \textit{TSB}_1
\end{array}
$$

Figure 9: Indicator translated to Linear Temporal Logic

$$
\begin{array}{l}
\quad \text{Regular-treatments}\#(;...) \wedge \langle \textit{environment assumption} \rangle \\
\rightarrow \langle \textit{indicator} \rangle
\end{array}
$$

Figure 10: Proof obligation with environment assumption

B.4 Verification

The formal verification of the indicator for the Jaundice guideline is a rather large case study in interactively verifying a temporal formula for a complex concurrent system. Verifying large systems interactively requires sophisticated tool support. Our experiences with the existing tool KIV, in which support for the verification of concurrent systems is rather new, are presented here.

In Section B.4.1 we define the property which is to be verified. Section B.4.2 shortly explains the proof strategy implemented in KIV. The following sections present our verification journal.

B.4.1 Proof obligation

We aim at verifying the indicator for the Regular-treatments plan. The resulting proof obligation is given in Figure 10: Regular-treatments implies the indicator property. The Regular-treatments plan is only a part of the Jaundice treatment. This plan does not measure TSB at all. However, this plan is executed in parallel to another plan which regularly draws measurements. If we want to verify the indicator for the Regular-treatments plan, we need to formulate restrictions for the environment.

An appropriate formula for the formulation of these restrictions, which we

$$
\begin{array}{l}
/^{*}\ \textit{environment assumption}\ ^{*}/ \\
\qquad \textit{pd}''\ .\text{under-phototherapy} \leftrightarrow \textit{pd}'\ .\text{under-phototherapy} \\
\quad \wedge\ \textit{time}'' = \textit{time'} + 1 \\
\quad \wedge\ /^{*}\ \textit{TSB measurements at most every 12 hours}\ ^{*}/ \\
\qquad\quad \textit{pd}''\ .\text{tsb} \neq \textit{pd}'\ .\text{tsb} \\
\qquad \rightarrow \exists\ T_0.\ T_0 = \textit{time}' \wedge \bullet \Box\ \textit{pd}''\ .\text{tsb} \neq \textit{pd}'\ .\text{tsb} \rightarrow 12 \leq \textit{time}' - T_0 \\
\textbf{unless laststep}
\end{array}
$$

Figure 11: Environment assumption

will call environment assumption, is given in Figure 11. The environment is expressed as a relation between primed and doubly primed variables v' and v''. Our formulae assumes that the environment:

- does not modify the *pd* .under-phototherapy flag (which indicates whether the patient is under phototherapy),
- always increments time, and
- measures TSB at most every 12 hours.

B.4.2 Proof strategy

The recipe to verify concurrent systems in KIV incorporates the following ingredients:

1. Symbolic Execution

   Parallel programs are executed step by step. The property is proven for the first step, a step of the program is then executed and afterwards verification continues with the rest of the program. The technique of symbolic execution can also be applied to temporal formulae.

2. Induction

   Symbolic execution may not terminate, because programs may contain loops. For executing loops, induction is necessary. Certain operators of temporal formulae also define a loop and ctherefore induction must be applied.

3. Modularity

   Large programs cannot be executed in a monolithic block, because proofs would be too large. The proof strategy in KIV allows to abstract sub programs by simpler properties. This is done by defining lemmas.

4. Sequencing

   Executing parallel programs may lead to a large number of cases. Often, in several cases the same parallel program remains to be executed and only some intermediate steps are different. Sequencing can be applied to unite these cases. Sequencing is currently not properly implemented in KIV.

B.4.3 Modular approach (first attempt)

As the overall behaviour of plan Regular-treatments is rather complicated, our first approach was to separately verify properties for its five sub plans. We would then use these properties as lemmas for verifying Regular-treatments.

The most difficult part in this approach is to come up with the appropriate lemmas. On the one hand they should abstract from the behaviour of the sub plans to allow for a simple proof for Regular-treatments. On the other hand they must not omit important aspects of the sub plans for the overall property to hold.

As we were trying to verify that the number of TSB measurements is restricted, it should be important to know that the sub plans do not measure TSB at all. For the Observation plan, this gives rise to the following lemma:

$$\text{Observation}\#(...) \quad \rightarrow \quad ((pd'.\text{tsb} = pd.\text{tsb}\ \textbf{unless laststep}) \qquad (1)$$

We can read it as follows. If we execute Observation, the history of TSB measurements stored within the patient record *pd* will not be changed, i.e. the history in the next state pd' .tsb will always be equal to the history in the current state *pd* .tsb. Similar lemmas can be formulated for all the other sub plans.

Proving these lemmas was rather straightforward and hence their proofs are omitted here. It should be mentioned though that the verification of these lemmas gave rise to the implementation of a number of heuristics with which the proofs were finally fully automatic. Although the proofs required induction, the necessary invariants were very simple, which made a heuristic approach possible.

Using these lemmas, we attempted to prove the indicator for Regular-treatments. The lemmas guarantee that all of the sub plans do not modify the history of TSB measurements. Thus, the whole body of Regular-treatments does not measure TSB, leaving us with a proof obligation of the following kind:

/ Remaining part of Regular-treatments */*
break
  break
    pd' .tsb = *pd* .tsb **unless laststep**
  if rt-continuation(...)
if rt-abort(...);
if rt-abort(...) **then** *rt-state* := aborted **else** *rt-state* := completed
$\wedge$ ⟨*environment assumption*⟩
$\wedge$ ...
$\rightarrow$ ⟨*indicator*⟩

We know that the environment measures TSB at most every 12 hours. We have to prove that –after phototherapy is discontinued– no more than one measurement is taken. Using the lemmas, we have abstracted the body of the plan such that the plans in the body do not measure TSB. But, phototherapy can be discontinued at any time and the body could keep on running forever. Thus, the environment would have time to measure TSB more than once and as result the property would be violated.

It follows that we need to know more about the behaviour of the sub plans. Especially how long it takes to complete the execution of the body after phototherapy has been discontinued. Plan Observation is a possible follow-up after a phototherapy has been tried. If Observation takes longer than 24 hours to complete, the property is violated. Thus, the lemma for Observation must also contain information about time. However, the Observation plan does not contain any restriction on the execution time and consequently the plan is allowed to run longer.

After this first proof attempt, our claim was that the Jaundice guideline does not satisfy the MAJIC indicator. A counter example is the trace where we apply phototherapy once and then we do observation. In this case Observation takes more than 24 hours to complete allowing the environment to measure TSB twice.

B.4.4 Assumptions

The original Jaundice guideline does not satisfy the MAJIC indicator. However, our proof attempt gave rise to certain assumptions under which the indicator might hold. From the medical point of view, it could be very helpful if, in addition to the counter examples that violate the property, we could describe the circumstances under which the property is satisfied.

Our initial assumption is that Observation is completed within less than 24 hours. However, if we examine more possible traces of execution, we discover additional counter examples. For example, if phototherapy has decreased the level of TSB to observation level, we would discontinue phototherapy and start Observation. If, during Observation, the TSB level again increases, another round of phototherapy would be necessary which would take more time and also require additional TSB measurements.

After considering a number of execution traces, we finally arrived to the following three assumptions:

(1) If phototherapy is discontinued, measured TSB will be of observation level.

(2) Observation will run for less than 24 hours.

(3) After phototherapy has been discontinued, the measured TSB will remain at observation level during the following Observation.

These assumptions must be added to the behaviour of the Jaundice guideline. It is up to medical experts to judge if these assumptions are too strong. The experts must decide whether it is enough that the indicator only holds in the assumed cases or whether the guideline must be improved such that the indicator holds in other cases as well.

In order to continue with the proof, we added the assumptions as additional axioms to the KIV formalisation of Jaundice. For example, assumption (2) is formalised as follows:

$$\begin{aligned} &\text{prescribe-observation\#(...)} \\ \rightarrow\ &\exists\ \mathit{time}_0.\ \mathit{time}_0 = \mathit{time} \wedge \Box\ (\mathbf{laststep} \rightarrow \mathit{time} - \mathit{time}_0 < 24) \end{aligned}$$

We can read this as follows. If we prescribe observation, we assume that the difference between the starting time time_0 and the time *time* in the last step is less than 24 hours.

Counter examples gave rise to assumptions. These assumptions could be helpful for medical experts to decide if guideline needs to be improved.

B.4.5 Modular approach (second attempt)

Our first attempt to prove the indicator for Regular-treatments failed, because the lemmas abstracting the behaviour of sub plans were too weak. Using the assumptions we were able to formulate stronger lemmas. The two assumptions (2) and (3) apply to the Observation plan. Therefore a stronger lemma for Observation should incorporate its old lemmaand both assumptions.

Again, the success of the overall proof will depend on the abstractions of sub plans. Does the abstraction contain enough information to complete the indicator proof? Discussing the outline of a possible proof led us to believe that this is

not the case. We still need more detailed information about the behaviour of sub plans. Among others, we need to know more about the *pd* .under-phototherapy flag within the patient record.

It is too difficult to guess a suitable abstraction of sub plan behaviour before trying the main proof. On the other hand, with the current tool support it is very time consuming to repeat the attempts on the main proof over and over again to gradually refine the lemmas. To automatically reuse proofs in KIV after lemmas have been changed is already possible for proofs in predicate logic and sequential programs. It would be very helpful if the reuse of proofs was also possible for temporal logic and parallel programs.

A more sophisticated method for coming up with lemmas/abstractions for sub plans is required. Also, tool support for the automatic reuse of proof attempts would be very helpful.

B.4.6 Nonmodular approach

As it was too difficult to guess strong enough lemmas for sub plans in advance, and too time consuming to retry the main proof several times, we decided to change our proof strategy and to try the main proof without abstracting the sub plan behaviour. An advantage of this approach is that, without abstractions, simple heuristics can be used to automatically execute the Regular-treatments plan. A disadvantage is that many more different cases have to be considered. For example, while executing the Observation sub plan, we always need to distinguish two cases: whether the sub plan aborts or continues to run.

As it could be be expected, automatic execution alone resulted in too many different cases –in fact the number of cases increased exponentially during execution. We were not able to follow all of the cases, but decided to finish the proof for some of the most interesting cases first. This led us to the following strategy:

1. use heuristics to symbolically execute plan automatically
2. keep track of case distinctions
3. follow most interesting branches

Among others, we have examined the case where Phototherapy-intensive is applied and completes successfully. Assumption (1) guarantees that measured TSB is of observation level afterwards. Thus Observation is applied and other treatment options are rejected. With assumption (2) Observation terminates within 24 hours. It must complete successfully because of assumption (3). If Observation completes, Regular-treatments completes and the proof for this case is finished. As can be seen, all three assumptions were required to finish the proof.

We have proven the indicator for special cases, leaving us with a proof with 1400 steps, but also 120 open sub goals. How do we ever complete all the sub goals?

Symbolic execution for parallel programs without abstraction generates too many cases and only selected cases can be proven.

B.4.7 Generalisation

The partial proof for the indicator contains 120 open sub goals, but also a number of completed cases. A complete proof for one case contains important information about the formulae needed to actually prove this case. This gives rise to generalisation.

As an example, we can consider the following sub goal:

$$
\begin{array}{l}
\quad /\!* \textit{ system } *\!/ \\
\quad \textbf{if } \text{rt-abort}(\ldots) \\
\quad \textbf{then } \textit{rt-state} := \text{aborted} \\
\quad \textbf{else } \textit{rt-state} := \text{completed} \\
\wedge\ /\!* \textit{ assumption (1) } *\!/ \\
\qquad \textit{pd} \text{ .under-phototherapy} \wedge \neg\ \textit{pd}' \text{ .under-phototherapy} \\
\qquad \rightarrow \text{get-bilirubin}(\textit{pd}' \text{ .tsb}) = \text{observation} \\
\quad \textbf{unless laststep} \\
\wedge\ /\!* \textit{ environment assumption } *\!/ \\
\qquad \textit{pd}'' \text{ .under-phototherapy} \leftrightarrow \textit{pd}' \text{ .under-phototherapy} \\
\qquad \wedge\ \textit{time}'' = \textit{time'} + 1 \\
\qquad \wedge \quad \textit{pd}'' \text{ .tsb} \neq \textit{pd}' \text{ .tsb} \\
\qquad \rightarrow \exists\ T_0.\quad T_0 = \textit{time}' \\
\qquad\qquad \wedge \bullet\ \Box\ \textit{pd}'' \text{ .tsb} \neq \textit{pd}' \text{ .tsb} \rightarrow 12 \leq \textit{time}' - T_0 \\
\quad \textbf{unless laststep} \\
\wedge\ /\!* \textit{ history } *\!/ \\
\quad \textit{pd} \text{ .under-phototherapy} \leftrightarrow \textit{pd}_0 \text{ .under-phototherapy} \\
\quad \wedge\ \textit{time} = \textit{time}_0 + 1 \\
\quad \wedge\ \textit{pd} \text{ .tsb} = \textit{pd}_0 \text{ .tsb} \\
\quad \wedge\ \text{rt-filter}(\textit{pd}_0) \\
\quad \wedge\ \text{rt-abort}(\ldots) \\
\rightarrow\ /\!* \textit{ indicator } *\!/ \\
\Box \quad \textit{pd} \text{ .under-phototherapy} \wedge \neg\ \textit{pd}' \text{ .under-phototherapy} \\
\quad \rightarrow \quad \Box\ \textit{pd} \text{ .tsb} = \textit{tsb}_1 \wedge \textit{tsb}_1 \neq \textit{tsb}_0 \rightarrow \Box\ \textit{pd} \text{ .tsb} = \textit{tsb}_1 \\
\qquad \vee\ \textbf{laststep}
\end{array}
$$

This sub goal contains a number of formulae. They do not need to be fully understood here. A proof for this goal has been automatically constructed by heuristics. Analysing this proof reveals that only part of the formulae were actually needed to prove the goal. This information can be used to construct a more general lemma:

$$
\begin{array}{l}
\quad /\!* \textit{ system } *\!/ \\
\quad \textbf{if } \text{rt-abort}(\ldots) \\
\quad \textbf{then } \textit{rt-state} := \text{aborted} \\
\quad \textbf{else } \textit{rt-state} := \text{completed} \\
\rightarrow\ /\!* \textit{ simplified indicator } *\!/ \\
\quad \Box\ (\quad \textit{pd} \text{ .under-phototherapy} \wedge \neg\ \textit{pd}' \text{ .under-phototherapy} \\
\qquad \rightarrow \textbf{laststep})
\end{array}
$$

The essence of this case is that if only the last conditional of the Regular-treatments plan remains to be executed, then a simplified version of the property can be proven without any further assumptions. This lemma can be used to

prove the case from where it originates, but can also be applied to at least nine more cases.

With the same strategy, we have generated five lemmas in a short amount of time and used those to prove 40 out of 120 open goals.

Proven cases can be analysed to generate lemmas. These lemmas can be applied in many other cases. The analysis could be automated in the future.

B.5 Results

Verification is not straightforward, but an iterative process of proof attempts revealing errors. The proof attempts can be used to generate counter examples. This is what we have done for the Jaundice indicator. Thus, the guideline of AAP does not comply with the MAJIC indicator. However, the proof attempts also suggested assumptions under which the indicator is satisfied. These assumptions have been given to medical experts for comments. It seems that the assumptions indeed capture the normal case, i.e. for most of the treated newborns the assumptions are valid. Therefore, in the normal case, the indicator is satisfied and TSB is measured only once after phototherapy is discontinued. In exceptional cases, it might not be desirable to enforce the indicator, because additional TSB measurements are probably required. The assumptions can be used then to explicitly define exceptional cases where treatment is not as usual. We believe that the quality of the original guideline could be improved by mentioning that it was not designed to satisfy the indicator in all cases. It could also specify that, in exceptional cases –which can be obtained from the assumptions– the indicator may be violated.

Formal verification of a property revealed an error and provided a counter example. Proof attempts suggested assumptions under which the property is valid. Assumptions can be used to explicitly define the exceptional cases under which satisfying the property may not be possible. Our claim is that adding these elements to the original guideline can improve its quality.

B.6 Conclusion

In this case study we have shown how formal verification can be used to analyse and improve medical guidelines. If guidelines are developed with certain goals in mind, formal methods can be used to verify whether the guideline achieves the goals. If not, counter examples can be provided and assumptions necessary for the intention to hold can be generated. Both can be used to improve the quality of the original guideline.

Sophisticated tool support is essential for the verification task. Currently the KIV tool is not powerful enough to do verification of large concurrent systems. It turned out that formulating lemmas for a modular proof is difficult. During the main proof, the lemmas abstracting the behaviour of sub plans often need to be changed resulting in repetition of proofs. During our case study, the following two additional ingredients to the proof strategy of KIV were dearly missed:

- Reuse of proof attempts

  After errors have been corrected, the KIV system currently allows for the automatic reuse of proof attempts for sequential programs and predicate

logic. This feature also needs to be implemented for parallel programs and temporal logic.

- Generalisation

  It should be possible to automatically analyse completed proof branches and to extract those parts of the formulae which were really needed to prove the sub goal. This would give rise to lemmas which could be applied to similar situations. This feature is very important, because the current proof strategy for verifying concurrent systems results in many similar cases.

References

[1] AAP. American Academy of Pediatrics, Provisional Committee for Quality Improvement and Subcommittee on Hyperbilirubinemia. Practice parameter: management of hyperbilirubinemia in the healthy term newborn. *Pediatrics*, 94:558–565, 1994. See also `http://www.aap.org/policy/hyperb.htm`.

[2] Deliverable D3' – KIV formalisation of reference protocols. www.protocure.org, May 2002.

[3] MAJIC Steering Committee Meets. *MAJIC Newsletter*, 1(2), October 1998.

[4] M. Marcos, M. Balser, A. ten Teije, and van Harmelen. From Informal Knowledge to Formal Logic: A Realistic Case Study in Medical Protocols. In *Proc. of the 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW-2002)*, pages 49–64, Sigüenza, Spain, October 2002. Springer-Verlag.

DELIVERABLES TABLE

Project Number:	***IST–2001-33049***
Project Acronym:	**PROTOCURE**
Title:	***Improving medical protocols by formal methods***

Del. No.	Revision	Title	Type[1]	Classification[2]	Due Date	Issue Date
D0	**1**	**Project presentation**	**O***	**Pub.**	**28/2/2002**	**26/2/2002**
D1	**1**	**Reference protocols and their Asbru models**	**R**	**Pub.**	**28/2/2002**	**28/2/2002**
D2	**1**	**Formal semantics of main Asbru elements**	**R**	**Pub.**	**31/3/2002**	**28/3/2002**
D3	**1**	**Desirable/required properties of main Asbru elements**	**R**	**Pub.**	**31/3/2002**	**3/4/2002**
D3'		**KIV formalisation**	**D**	**Pub.**	**31/5/2002**	**31/5/2002**
D4		**Verification of properties on the reference protocols**	**R**	**Pub.**	**30/9/2002**	**9/10/2002**
D5		**Evaluation of results**	**R**	**Pub.**	**30/11/2002**	

1 *R: Report; D: Demonstrator; S: Software; W: Workshop; O: Other – Specify in footnote*

2 *Int.: Internal circulation within project (and Commission Project Officer + reviewers if requested)*
Rest.: Restricted circulation list (specify in footnote) and Commission SO + reviewers only
IST: Circulation within IST Programme participants
FP5: Circulation within Framework Programme participants
Pub.: Public document

* Report, webpage.