

DELIVERABLE SUMMARY SHEET

Project Number: *IST-2001-33049*

Project Acronym: **PROTOCOLURE**

Title: *Improving medical protocols by formal methods*

Deliverable N°: D5

Due date: 30/11/2002

Delivery Date: 2/12/2002

Short Description:

This document contains the evaluation results of the Protocolure project. This deliverable consists of two parts. The first part is the *evaluation by medical experts*. It describes the benefits of quality improvement of protocols by formal methods, the evaluation of relevance of detected problems and the potentials of our approach. The second part is the *evaluation by the project team*. It concerns the feasibility of quality improvement of protocols by formal methods. The whole formalisation and verification process is evaluated. We describe the results, the lessons learned and the open issues including challenging issues.

Partners owning: *all*

Partners contributed: *all*

Made available to: Pubic

D5 Evaluation results

Version 1.0

December 2, 2002

Abstract

This deliverable consists of two parts. The first part is a description of the evaluation of our results by medical people. This document is based on interviews with of specialists on the subject of the selected medical protocols. The second part is an evaluation of our results by the project team themselves. In this part we evaluate our approach of using formal methods for improving medical protocols. We pay attention to the results, the lessons that we have learned, and the open issues (challenging).

Contents

Evaluation results, medical part

Evaluation results. technical part

Evaluation of results, medical part

Version 1.0
December 2, 2002

Contents

<u>Introduction</u>	1
<u>Methods</u>	2
<u>Results</u>	4
<u>Conclusion</u>	5
<u>Appendix 1: Background paper and glossary</u>	6
<u>Protocure background paper</u>	6
<u>Protocure Glossary</u>	7
<u>Appendix 2: Summaries of the three interviews</u>	9

Introduction

In this part of D5 we describe the evaluation by medical experts of the relevance of detected flaws and the use of formal methods for quality improvement of guidelines and protocols.

Medical guidelines are systematically developed statements to assist the practitioner and patient in making decisions about appropriate healthcare for specific conditions. The term protocol is used for more specific versions of a guideline, often incorporating specific knowledge to fill in the gaps in the general guideline with local details. A guideline or protocol describes the optimal care for patients, and therefore if applied well improves the quality of patient care.

The evidence based guideline development process is a structured and systematic process: systematic search and selection of literature, critical appraisal and grading of the literature are important steps. Transparency of every step is important. Before 1995 guidelines were constructed in a less systematic way and were therefore less transparent. These guidelines are often called consensus guidelines.

Medical guidelines assist practitioners not only to perform the correct actions, but also to perform the actions correct. Quality has always been inherent in the actions of professionals, and is included in three essential questions: 1) What is the best care?; 2) Does a professional practice what he knows?; 3) Does a professional measure what he practice? Guidelines are providing tools to answer the first question. A guideline is therefore only one component of the quality system, which aims to continuously improve the quality of patient care. A coherent quality system consists of a continuous cycle of measurement – improvement – redesign and consolidation of care processes. To measure the compliance to the guideline and to evaluate the outcome of patient care, indicators can be formulated. Indicators are quantitative measures for monitoring clinical care to improve performance and quality.

Except for the direction to design more coherent quality systems in which guidelines are embedded, there are some other future challenges. An important one is to integrate guidelines more in the daily work-processes of the practitioners by using computerised support. To make this possible guidelines must be formalised in such a way that this becomes possible. Another development is that guidelines and protocols will be updated on a more continuous basis, so called “living guidelines”. Our aim is to develop guidelines which present up-to-date and state-of-the-art knowledge to the practitioners. Guidelines must be structured in a more modular way, so in case of modifications of one part of the guideline it is not necessary to revise the whole document.

Guidelines and protocols need to be of good quality. As stated by Basinski (1995): “Although few businesses would risk the failure of full-scale product releases without prior testing and test marketing, these preliminary activities have been largely ignored in most guideline development endeavours.”

At present, informal instruments such as AGREE¹ are used to assess quality, but these instruments are more focused on the methodology of the guideline development process than on a guideline's scientific content. Guidelines are complex documents. As a result, guidelines can be ambiguous, incomplete and even inconsistent. It is a precise and time consuming work to detect flaws in guidelines. The guidelines are developed and read by practitioners with their particular background knowledge, as a result they fail to notice even serious flaws. Therefore formal methods can be feasible to detect flaws in guidelines and to improve the quality of guidelines and protocols. To investigate this the Protocure project was started.

This document describes the opinion of three practitioners on the relevance of the detected flaws by using formal methods. They also evaluated the approach with formal methods. These practitioners were specialised in the field of the guideline subjects: jaundice and diabetes.

Methods

To obtain the opinion of medical experts on the use of formal methods to detect flaws in guidelines/protocols and the seriousness of the flaws for patient care, structured interviews were taken. The interviews were held between a Protocure member and an expert on jaundice or diabetes. The experts and their interviewers are mentioned in table 1. The diabetes expert judged the flaws detected in the diabetes guideline, and the pediatricians judged the flaws detected in the jaundice guideline.

Table 1. Experts and interviewers

Name expert	Specialism	Institute expert	Interviewer	Institute interviewer
Dr. F. Storms	Diabetes expert	Mesos Medical Centre, Utrecht; Diabetes Clinic Bilthoven	Ms. Dr. A. ten Teije	Vrije Universiteit Amsterdam
Prof. Dr. M. van de Bor	Pediatrician	University Medical Centre Nijmegen, Department of Pediatrics	Dr. P. Lucas	University of Aberdeen
Prof. Dr. C. Popow	Pediatrician	University Medical Centre Vienna	DI A. Seyfang	Technische Univesitaet Wien

Interviewing procedure

The interview took about two hours. Before the interview the expert received several documents. A background document for better understanding of the project was provided in advance (see appendix 1). An accompanying glossary explained the difficult terms used in the project and the background document (see appendix 1). Next to this they received the guideline on jaundice or diabetes and a questionnaire just to get an idea of the questions used in the interview.

The interviewer and the expert reviewed the questions one by one. Each answer was extensively discussed with the expert and the discussions were summarised by the interviewer. The summaries are enclosed in appendix 2.

Contents of the interview

The interview contained three aspects:

1. Opinion of experts of the quality of the guideline (diabetes or jaundice)
2. Relevance of the flaws found in the guidelines: criticality and necessity of formal methods
3. Relevance of the approach

¹ The purpose of the Appraisal of Guidelines for Research and Evaluation (AGREE) Instrument is to provide a framework for assessing the quality of clinical practice guidelines. The AGREE instrument assesses both the quality of reporting, and the quality of some aspects of recommendations. It provides an assessment of the predicted validity of a guideline, that is the likelihood that will achieve its intended outcome. The AGREE instrument consist of 23 key items organised in six domains. Each domain is intended to capture a separate dimension of guideline quality: 1) scope and purpose; 2) stakeholder involvement; 3) rigour of development; 4) clarity and presentation; 5) applicability; 6) editorial independence. (www.agreecollaboration.org)

Each aspect was translated into several questions:

Aspect 1: quality of the guideline:

- Are you familiar with the guideline on jaundice/diabetes?
- What do think about the quality (presentation and content) of the guideline?

Aspect 2: relevance of the detected flaws

- Consider the following list of uncovered flaws. Indicate for each flaw separately how serious the flaw is in your opinion (4 point scale: extremely serious vs. not serious).
- Would you regard the computer supported systematic analysis of flaws an improvement over current practice?

Aspect 3: Relevance of the approach

- Please mention three pros and cons of our approach
- What is your opinion about the general methodology (formalisation of protocol and verification of properties of the protocol)?
- How would you judge the potentials of this approach?

Selection of flaws

Not all the flaws found in the project were reviewed by the experts. A selection was made during a meeting of the Protocure partners. We selected a diversity of flaws detected by the formal approach.

List of used flaws:

Jaundice

- *incompleteness*: There is no definition for the actual rate TSB increase which should be considered a "rapid TSB increase" but this term is used in the guideline.
- *inconsistency*: It is not clear, whether clinically jaundiced children of less 24 hours are considered healthy newborns or not.
- *inconsistency*: Signs for an underlying disease given in three lists (table, diagram, and main text) and these lists are not consistent.
- *termination*: for jaundice patients who are no emergency cases, observing bilirubin levels goes on for ever (it is not explicitly stated when this can stop).
- the verification of two intentions revealed noteworthy details
 - photo-therapy-intensive: treatment intentions are only satisfied when a treatment plan is actually carried out and not stopped before the final aims are achieved (this was a result of application of the formal proof methods)
 - MAJIC indicator no 7: no more than one serum bilirubin level drawn after phototherapy is discontinued (this indicator holds only under certain assumptions)
- *Incompleteness*: unused test-results: child blood type (ABO, Rh) and Coomb's test

Diabetes

- *Incompleteness*: unused test-results of creatinin test
- *Incompleteness*: unused requested patient info:
 - alcohol consumption;
 - physical exercise habits
- *Incompleteness*: unspecified reasons for alternative decisions: different insuline treatments
- *Redundancy*: redundant data-requests: repeated questions about patient weight
- *Correctness*: correct doses: morning/evening dose distribution of 1/3-2/3
- *Correctness and safety*: adjust therapy at glucose levels < 3.5 mmol/l and > 8 mmol/l
- *Safety*: Safe dose increases: safe dose increase to avoid treatment overshoot
- *Safety*: Safe therapy: avoid combination of all three oral medication

The pediatricians judged the flaws in the jaundice guideline and the diabetes expert judged flaws in the diabetes guideline.

Results

1. Quality of the guidelines

The experts were familiar with the guidelines used in the project. The jaundice experts considered the quality of the jaundice guideline poor to acceptable. The main reason for this was that the guideline is outdated, it was published in 1994. It does not mention a new treatment option. Also is the scientific evidence of the guideline not always sound. For example, studies investigating the consequence of haemolysis in newborns indicate that the dangers for kernicterus of the levels of serum bilirubin mentioned in the guideline may have been overestimated.

The quality of the diabetes guideline was good according to the diabetes expert. It is based on literature and it is done very well. However he mentioned on forehand some gaps in this guideline: reference to the second line is not specified, it does not pay attention to “over treatment” and the insulin part is ‘problematic’.

2. Relevance of the detected flaws

Most of the flaws detected in the jaundice guideline were not considered serious by the experts from a clinical point of view. They had the opinion that most clinicians have enough background knowledge on the disease, but it makes the guideline less suitable for inexperienced doctors. The experts considered most flaws not as a danger for the patient.

However one expert, also involved in guideline development, stated that from a point of view of guideline developers some of the flaws are serious as it indicates lack of care in designing the guideline and the flaws should have been avoided.

Half of the flaws in the diabetes guideline were not considered very serious by the diabetes expert. The remainder of the flaws in this guideline he judged as very serious and even one extremely serious! For example an extremely serious flaw was that the guideline does not pay attention to over treatment. The too low value of glucose value (≤ 3 mmol) is not in the table in the guideline.

In general the three experts regarded the computer supported systematic analysis of flaws as an improvement. For experienced clinicians the computer supported flaw detection will have little consequences for clinical practice. However, for guideline developers it may be a very good method to improve the quality of clinical guidelines. Because of the background knowledge, the clinicians fail to see the flaws when reading the guideline. One of the experts mentioned that especially the logic of the guideline will improve. The method shows that serious flaws can be found and one does not want to oversee these flaws.

3. Relevance of the approach

Pros and cons of the approach

The experts mentioned the following pros and cons of the approach:

Pros:

- The dialog between the computer scientist and physicians. It forces the clinician to define his/her position exactly.
- Pointing out flaws in guidelines.
- Improvement of guidelines through this interaction process.
- Useful for detection of pitfalls and ambiguities.
- May give rise to production of better definitions.
- More straightforward and complete protocols/guidelines.
- Find gaps in the protocol. However it is an illusion to find them all.
- Formal approach is without competition of medical experts from first or second line.
- Necessary first step for more ICT support with respect to protocols. For instance an advice system, critiquing system.
- Get more insight in the protocol.

Cons:

- It is a very formalistic approach.
- Clinical daily practice is changing very fast and the reaction to changes takes very long. This is a general disadvantage of guidelines.
- Strict protocols/guidelines are violated by clinicians more easily.
- Strict protocols/guidelines may give rise to 'cookbook medicine'.
- Too much detail, and too little emphasis on medically relevant issues.
- Medical knowledge is complex. For instance for diabetes it is not possible to cover diabetes for 100% in a guideline. So there are always exceptions.
- Medical experts might see it as critique.
- The verification of protocols/guidelines is labour intensive.
- The transfer of the results (flaws) is labour intensive too.
- There is a need for non-medical people in developing guidelines/protocols.

One expert gave additional remarks. First, he stated that the method is not transferable to medical people. This is also considered as a pro, since another way of thinking enables finding flaws. Second, the end-point of the approach is not the list of flaws, but the discussion of these points. This discussion should have some effects!

Opinion about the general methodology

Use of formal methods during the process of guideline development may improve guidelines considerably. One expert stated that it forces one to be explicit about intentions of (parts of) guidelines. This is important, especially for patient goals. It enables to relate guideline to indicators (goals). It is important to monitor indicators both from individual patient and from patient groups.

Potentials

The experts judged the potential of this approach, i.e. the use of formal methods, especially promising for guideline developers. It is also an important first step to make further ICT support for guidelines possible.

Conclusion

Flaws in guidelines can be found by formal methods. Most flaws are not really a danger to patient care, because of the background knowledge of the experienced doctor. However from the point of view of guideline developers some of the flaws are serious as it indicates lack of care in designing the guideline and the flaws should have been avoided. Therefore the experts in this evaluation considered the use of formal methods especially useful for guideline developers.

To incorporate guidelines or protocols in ICT systems, the guidelines or protocols must be developed in more structured way. Formal methods can be of help here. This is an important first step to enhance further ICT support for guidelines and protocols.

Appendix 1: Background paper and glossary

Protocure background paper

Terminology in this document is clarified in a accompanying glossary.

Protocure is a project within the Information Society Technology (IST) Programme from the European Union. The project's full title is: 'Improving medical protocols by *formal methods*'. *Protocure* is a one-year project, starting December 2001 and has to be finished by the end of November 2002.

Participating organisations are:

Vrije Universiteit, Faculty of Sciences, division of mathematics and computerscience.

Technische Universitaet Wien, Institut fuer Softwaretechnik und Interaktive Systeme.

Universitaet Augsburg, Institut fuer Informatik, Lehrstuhl fuer Softwaretechnik und Programmiersprachen.

Dutch Institute for Healthcare Improvement CBO, department of Guideline Development.

University Court of the University of Aberdeen, department of Computing Science.

Problem Statement

Clinical guidelines are systematically developed statements to assist practitioner and patient decisions about appropriate healthcare for specific clinical circumstances. The recommendations in guidelines are based on scientific evidence (*Evidence Based Medicine (EBM)*). The term *protocol* is used for more specific versions of a guideline, often incorporating specific knowledge to fill in the gaps in the general guideline with local details.

Guidelines and protocols have to be of good quality, currently quality is assessed by instruments like *AGREE* (informal methods). These instruments assess more the methodology of the guideline development process than the scientific content. As a result most guidelines are still ambiguous, incomplete and inconsistent.

Objective of this project is the evaluation of the feasibility of the use of formal methods for quality improvement of guidelines or protocols.

Work description

Two guidelines (jaundice in newborns and diabetes mellitus type 2) were gradually transformed into a formal representation. Starting from the text version, the guidelines were first modeled in the language *ASBRU*, an intermediate between informal guidelines and the purely formal method used by *KIV*. After that the guidelines were translated into *KIV*. In parallel, a number of *properties* were determined in the guidelines. Properties are features of the guideline that describe the critical points of care. The actual proof of this properties in *KIV* will serve to assess the feasibility of using formal *verification* for guideline/protocol improvement. The essence is that the computer proves if the recommendations in guidelines satisfy the properties stated by experts.

Expected results

Measure of success is to uncover a certain amount of flaws in the guidelines. This uncovering of flaws is executed in the formalisation as well in the verification process.

The evaluation of the project consists of two parts:

Part 1: Evaluation by medical experts of the relevance of problems detected in guidelines/protocols.

Part 2: Evaluation by *Protocure* partners of the feasibility of the approach.

Part 1: Evaluation by medical experts

The evaluation in part 1 is done by separate interviews between *Protocure* members and the medical experts. The interview contains the following aspects:

- Opinion of experts of quality of the guideline:
 - Management of Hyperbilirubinemia in the Healthy Term Newborn (October 1994). American Academy of Pediatrics. (www.aap.org/policy/hyperb.htm)
 - NHG-standaard Diabetes Mellitus type 2 (2001). Dutch College of General Practitioners. (<http://nhg.artsennet.nl/standaarden/M01/std.htm>)
- Relevance of the selected properties
- Relevance of the flaws found in the guidelines: criticality and necessity of formal methods
- Relevance of the approach

Protocure Glossary

Protocure

Project acronym.

Clinical guidelines

Clinical guidelines are systematically developed statements to assist practitioner and patient decisions about appropriate healthcare for specific clinical circumstances.

Protocols

A rule of procedure or set of instructions containing conditional logic for solving a problem or accomplishing a task.

Evidence based medicine (EBM)

EBM is the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients. The practice of EBM means integrating individual clinical expertise with the best available external clinical evidence from systematic research.

AGREE-instrument

The purpose of the Appraisal of Guidelines Research & Evaluation (AGREE) Instrument is to provide a framework for assessing the quality of clinical practice guidelines. The AGREE instrument assesses both the quality of reporting, and the quality of some aspects of recommendations. It provides an assessment of the predicted validity of a guideline, that is the likelihood that it will achieve its intended outcome.

Formal methods

Formal methods are techniques to develop systems with high reliability, for safety-critical applications. Formal methods provide means to define a formal model (*formalisation*) and to formally verify properties (*verification*). Formal methods are based on formal semantics, which exactly define the intended meaning.

ASBRU

Asbru is a semi-formal language for describing medical protocols in a computer readable manner. It represents protocols as hierarchical plans of steps for treatment and diagnosis. Important aspects of Asbru are that it allows to specify temporal aspects of these steps and that each step can have associated intentions.

KIV

Tool with interactive theorem prover for the formal development of safety-critical systems. It supports modular specification, validation and verification. Verification is automatic to a large extent.

Verification

Verifying with mathematical rigour that a given system satisfies certain properties. A logical calculus ensures the correctness of verification. Verification requires tool support.

Formalisation

The process of defining a formal model (specification or implementation) for a given system description. During formalisation, ambiguities in the system description need to be removed and details need to be added. The resulting formal model exactly defines the system's behaviour.

Properties

Sentences describing some aspect of a system's behaviour. Properties can also be formalised. Verification is used to check, if a system satisfies certain properties.

Intentions

Intentions are high-level goals of a plan. These intentions should hold during the execution of plans/actions or after finishing a plan/action. They serve various tasks. In verification, they are the

basis for the formulation of properties, i.e., one important aspect is to verify that intentions are fulfilled by the plans implementing them.

Indicators

A quantitative measure for monitoring clinical care to improve performance and quality.

Flaws

Flaws are indications of potential errors in a protocol. They are discovered during modelling the protocol (first in Asbru, then in KIV), or during the attempt to prove that some expected properties of the protocol (e.g. intentions, indicators) do indeed hold.

Appendix 2: Summaries of the three interviews

Protocure Expert Questionnaire

(Dr. F. Storm, diabetes expert)

1. Are you familiar with the guideline on diabetes?

Yes	No
-----	----

Explanation:

The NHG diabetes protocol is the most detailed diabetes protocol. NHG diabetes protocol has more detail than the CBO diabetes version. These protocols have different focus. The NHG protocol is more practical/process oriented where as the CBO version has a more theoretical approach. In general the CBO protocols are more general, and tell less about the process (how to do things) than NHG-protocol.

Special to the NHG standard is that a team with medical people from the first line (GP's) and medical people from the second line (specialists) had developed this protocol (Diabetes is a disease that is treated in both lines).

2. What do you think about the quality (presentation and content) of the guideline?

Poor	Acceptable	Good	Excellent
------	------------	------	-----------

Explanation:

The approach that is used for writing this protocol is ok. It is based on the literature and it is done very well.

There are some gaps in this protocol:

- The reference to the second line is not very good included in the protocol. This is a sensitive point, since the thought of "GP's can not do this or that". In general are the relations between first and second line medical people not very good. For this purpose the reference to the second line is on purpose not very well specified in the protocol.*
- A real defect: the protocol does not pay attention to "over treatment"*
- Another real defect: the insulin part of the protocol is problematic.*

In total the content is acceptable and the presentation (plastic card, long version) too.

3. Consider the following list of uncovered flaws. Indicate for each flaw separately how serious the flaw is in your opinion.

- *Incompleteness: **unused test-result of creatinin test** (fourth bullet risk invent.)*

Extremely serious	Very serious	Serious	Not serious
-------------------	--------------	---------	-------------

There are two types of creatinin test: (1) in the blood and (2) in the urine (this in combination with the albumin). The blood test of creatinin is a test for the kidney functioning. If the creatinin is high then there are actions necessary: (1) to determine the risk, and (2) possibly second line reference (dialyse).

Considering the creatinin blood test, this test is used at the end of the protocol.

("consult & reference; serum creatinin > 2000 mmol/l or < 30 ml/min").

Concerning the amount of creatinin in the urine. It is right that this test is not used somewhere.

However there are two indicators before this one: protein in the urine, and high blood pressure, (lipide too). If these indicators are not good, then these will be treated. The creatinin increases only if both these indicators are bad. So, although the creatinin (urine) test is not really used, the “pre” indicators are. This means that the risk factor of a too high creatinin is covered by looking to these “pre” indicators. By the way a high creatinin is a serious problem.

- **Incompleteness: unused requested patient info: alcohol consumption, physical exercise habits** (second bullet risk invent.)

Extremely serious	Very serious	Serious	Not serious
--------------------------	---------------------	----------------	--------------------

This should be more explicit in the protocol. In the protocol is written “Give information about important aspects of DM and situations where actions are required.” The GP could say here something about alcohol consumption and physical exercise habits. For instance that alcohol influence the (decrease of) glucose level and risk on heart diseases. However, it would be better to write this more explicit in the protocol.

- **Incompleteness: unspecified reasons for alternative decisions, different insulin treatment** (insulin + oral, or only insulin)

Extremely serious	Very serious	Serious	Not serious
--------------------------	---------------------	----------------	--------------------

The intention is that first the insulin&drug treatment should be given, and if this does not work the treatment with only insulin.

The headers in the protocol are confusing, and actually they are not at the right place. The headers suggest a “or” between these two treatments. However it should be read as sequential steps.

This flaw is categorised as “serious” only for presentation reasons.

- **Redundancy: redundant data requests, repeated question about patient weight** (3-monthly and yearly control)

Extremely serious	Very serious	Serious	Not serious
--------------------------	---------------------	----------------	--------------------

Nobody will measure the weight twice.

- Verification of correctness of a treatment: **Correct doses: When using 2 doses per day of insulin for Diabetes treatment, a good practice to distribute the morning and evening doses according to the ratio 2/3-1/3.**

Extremely serious	Very serious	Serious	Not serious
--------------------------	---------------------	----------------	--------------------

It seems that the protocol designers have chosen for an efficient algorithm. Efficient in the sense of an easy algorithm. That means that during the increase of the drugs the ratio 2/3-1/3 is violated. This ratio 2/3-1/3 has less risk for the patient, and should be preferred. This property corresponds with the clinical practice.

(In the protocol at the end of insulin&drugs treatment is this morning/evening ratio almost 1/3-2/3!)

This flaw is categorised as very serious. The combination with taking the postprandial value increases the seriousness of this flaw.

- Verification of safety of a treatment: **Safe dose increases: In the Diabetes treatment it is not advisable to increment drug/insulin doses too quickly unless the patient is in a critical situation. The idea is avoiding the negative effects of treatment overshoot, i.e. hypoglycemia episodes. More precisely, the dose should be titrated (i.e. increased by smallest amount) in intervals of at least 2 to 4 weeks, unless the blood glucose is that high (e.g. > 15 mmol/l) that a faster pace of dose titration is needed.**

Extremely serious	Very serious	Serious	Not serious
--------------------------	---------------------	----------------	--------------------

Safety is very important, especially with respect of avoiding hypo's. It is better to go slowly to the optimal situation than rapidly. This because of the inconvenience for the patient (the so called "insulin blues") and because of the increase of risk.

- Property: **avoid combination of all three oral medication** (see after step 3)

Extremely serious	Very serious	Serious	Not serious
--------------------------	---------------------	----------------	--------------------

This is a sentence that is written in the protocol. It is a good advice. It would be useful to know whether this holds if the protocol is applied and medication is prescribed.

- Correct evaluation: **Blood glucose levels below 3.5 mmol/l (fasting and postprandial) are hypoglycaemic and not good. It may be necessary to adjust the therapy. Fasting blood glucose values above 8 mmol/l and postprandial glucose levels above 10 mmol/l are hyperglycaemic and not good.** (see table with blood glucose values, <3.5 mmol is trivially violated)

Extremely serious	Very serious	Serious	Not serious
--------------------------	---------------------	----------------	--------------------

Hypoglycaemic is the most urgent side effect of the treatment. This protocol does not pay attention to "over" treatment. That means it does not take into account whether the amount of drugs that the patient used is too high. This is in particular important in the insulin&drugs treatment, since there is the amount of drugs very high at the end.

Hypoglycaemic is very urgent, but does not occur very often. In the protocol is the extreme variant of hypoglycaemic covered, namely coma. In this case the protocol gives the advice of "find the cause of the coma". This cause can be for instance "over" treatment, no food, alcohol consumption. (These causes are not explicit mentioned in the protocol). The too low value of glucose value (≤ 3 mmol) is not in the table in the protocol.

It seems that if the patient stops with oral drugs, then the safety ratio of 2/3:1/3 of drug amounts is taken into account. The amount of drug is very low at the beginning, and may be this is the reason for the advice of a very rapidly increase of drugs. Notice that the patient goes from an amount of 60 + oral drugs to an amount of 20! The insulin part of this protocol is not very well developed. May be this is a sign that GP's are not used to work with insulin.

4. Would you regard the computer supported systematic analysis of flaws an improvement over current practice?

Yes. Especially the logic of the protocol will improve. Medical people read the protocol very quickly and then they miss some aspects. The method shows that serious flaws can be found. One does not want to miss these flaws.

In general the protocol developing team is only willing to change the protocol when there is a real changes. For instance, new drugs are available. A developing team is rather proud on its protocol, and they do not want critic.

For instance if we consider the previous version of diabetes and the current one. There are big difference in these protocol, for instance the whole insulin part is new, the protocol is much more process oriented (3-month/year controls), much more patient info.

What about applying our method only on parts of the protocol?

This seems very useful, for instance in diabetes the treatment part is very interesting, but the diagnosis part is not very interesting. However only looking to the insulin part is not an option, because then this part of the treatment is completely out its context.

The next step should be contacting the developing team and see what their reaction is. The expectation is rather negative/defensive.

5. Please mention three pros and cons of our approach.

Pros:

- 1. Find gaps in the protocol. However it is an illusion to find them all.*
- 2. Formal approach is without competition of medical experts from first or second line.*
- 3. Necessary first step for more ICT support with respect to protocols. For instance an advice system, critiquing system.*
- 4. Get more insight in the protocol*

Cons:

- 1. Medical knowledge is complex. For instance it is not possible to cover diabetes for 100% in a protocol. So there are always exceptions.*
- 2. Possibly, medical experts see it as critic.*
- 3. Labour intensive.*
- 4. The transfer of the results (flaws) is labour intensive too.*
- 5. There is a need for non-medical people in developing protocols.*

The method is not transferable to medical people. This is also considered as a pro, since another way of thinking enables finding flaws.

The end-point of the approach is not the list of flaw, but the discussion of these points. Our approach enables us to discuss these points of interests. This discussion should have some effects....

6. What is your opinion about the general methodology (formalisation of protocol and verification of properties of the protocol)?

The effort and the results depend on the quality of the protocol.

The NHG protocols are good candidates for this approach. They are very practical, from the same quality, and in general not complete.

It would be useful to apply the method to a more theoretical protocol (like the CBO's ones).

Forces one to be explicit about intentions of (part of) protocols. This is important, especially for patient's goals.

It enables to relate protocols to indicators (goals). It is important to monitor indicators both from individual patient and from patient groups.

7. How would you judge the potentials of this approach?

The method is now applicable, since flaws are found. These flaws should be communicated to the protocol developing teams.

The potentials of this approach depends a lot on political decisions...

Protocol designers will be very positive, especially those who really believe in protocols.

A commission that just has finished their protocol is proud, and will behave defensively.

Real use of a protocol is a difficult step. Medical people read them, and look whether they handle roughly in this way, and then the protocol disappear...

Important first step for further ICT support for protocols. Such system should be built on patient-data systems. However, at this moment these systems are not in a very good shape.

It would be nice whether the approach could derive a kind of indicators. Indicators (goals/measurements for patients or patient groups) will be very important in the future. The expectation is that medical care will be evaluated based on indicators.

Protocure Expert Questionnaire (Dr. C. Popow, pediatrician)

1. Are you familiar with the guideline on jaundice?

Yes	
-----	--

Explanation:

We are using a similar guideline. It was published 4 years ago by Paky in "Pädiatrie und Pädologie".

2. What do you think about the quality (presentation and content) of the guideline?

<i>Poor</i>			<i>Excellent</i>
-------------	--	--	------------------

Explanation:

Guidelines from this source are known to be excellent. But this one is outdated because it does not mention the treatment with immunoglobulin. On the other hand, a review from Cochrane (Issue 3, 2002) states that current studies are insufficient for recommending routine use of intravenous immunoglobulin.

3. Consider the following list of uncovered flaws. Indicate for each flaw separately how serious the flaw is in your opinion.

- **Incompleteness: *undefined rate of rapid TSB increase*** (item 1, 'Management of hyperbilirubinemia in the healthy term newborn by age' section)

			<i>Not serious</i>
--	--	--	--------------------

Explanation: This is well-defined to be 1 mg%/h (i.e., 1 mg/100ml/h) since the Seventies, e.g., Pollacek in "Kinderärztliche Notfälle" (pediatric emergencies), Georg Thieme Verlag, Stuttgart, 9th Edition, 1976.

- **Inconsistency: *are clinically jaundiced children of less than 24 hours healthy newborns?*** (line 1, item 1, section 'Management of hyperbilirubinemia in the healthy term newborn by age', caption Table 2, and heading flowchart are inconsistent)

			<i>Not serious</i>
--	--	--	--------------------

Explanation: These babies are otherwise healthy, so why not mention them in the guideline?

- **Inconsistency: *signs to rule-out underlying disease*** (table 1, item manifestations of the disease, versus box 2 in the flowchart, and item 7 'Evaluation' section text)

			<i>Not serious</i>
--	--	--	--------------------

Explanation: There is a reference to Table 1 in the text, and there should be one in the diagram, too, and there (in the table) the full list is given, so you can find it, if you want. It is only for space considerations that not the full list is repeated in the diagram and the text.

- **Termination:** 1 as flaw **I have no idea which part of the guideline is causing this problem**

			Not serious
--	--	--	--------------------

Explanation: Basically, the TSB-reading decreases after some time or it is a pathological case. Still if it is not mentioned in the guideline explicitly, this might be changed.

- **Intentions:**

- phototherapy-intensive: **time-annotation problem?**

Clarification of Asbru: An intention need not hold, when plan is not yet activated or is aborted.

			Not serious
--	--	--	--------------------

Explanation: It is clear, that a plan cannot guarantee anything if it is aborted.

- MAJIC indicator no 7: **no more than one serum bilirubin level drawn after phototherapy is discontinued** (more than one test consistent with guideline)

		Serious	
--	--	----------------	--

Explanation: Multiple blood samples are usually not necessary after phototherapy, if another value lies within a certain range. TSB usually raises by 1-2 mg/dl within 24 hours after the phototherapy and decreases afterwards. But this does not hold for the following cases:

1. after exchange transfusion,
2. if phototherapy was necessary before the 4th day of life,
3. if there is suspense for another underlying disease.

In addition, the exact time of the control (blood sample) should be defined. Normally, it is 24 hours after the previous one, but in case of expected problems (e.g., significant haemolysis), this is much earlier, e.g., after 6 hours.

- **Unused test results: child blood typing (ABO, Rh) and Coomb's test** (box 11 flowchart, item 1-3 text, section 'Evaluation')

			Not serious
--	--	--	--------------------

Explanation: These test give you information about the likely development of the TSB-readings. If a mother is, e.g., Rhesus-negative and the baby is Rhesus-positive, the blood cells of the baby cause the mother to produce anti-bodies against the babies red blood cells (some of) which go back to the babies body and kill red blood cells there. The baby starts producing additional red blood cells, but if they are not sufficient, the baby becomes anaemic, because there are not enough red blood cells to destroy the oxygen in the blood. After birth the mother's anti-bodies still sit on the red blood cell of the baby but are inactive and in the next days they are removed by the spleen.

The mentioned tests give information on which of the incompatibility is present and how many anti-bodies are present in the baby etc. There are many different cases of incompatibility, the one with the Rhesus-factor is only the simplest.

So this lack of information in the guideline is not serious, since the guideline alone is not enough anyway and you know these things, if you are a physician.

4. Would you regard the computer supported systematic analysis of flaws an improvement over current practice?

Of course! Because of the background knowledge, you skip flaws when reading a guideline, but the educational value of such analysis is very high. Afterwards, you can say: "We are still confused, but at a higher level."

5. Please mention three pros and cons of our approach.

Pros:

1. The dialog between computer scientists and physicians. It forces the clinician to define his/her position exactly.
2. Pointing out flaws in guidelines.
3. Improvement of guidelines through this interaction process.

Cons:

1. The labour involved.
2. It is a very formalistic approach.
3. Clinical daily practice is changing very fast and the reaction to changes takes long. This is a general disadvantage of guidelines.

6. What is your opinion about the general methodology (formalisation of protocol and verification of properties of the protocol)?

This is a very reasonable if not the only possible approach. I cannot say anything about the formal verification, but the evaluation of guidelines in daily practice is very important.

Verification/evaluation shows limits of guidelines, perfect ones are not possible. The knowledge doubles every 10 years [invalidating the existing knowledge] – still this [our work] is very important.

Note: text in [] inserted by interviewer.

7. How would you judge the potentials of this approach?

Great! Being on of the parents of it, I find it great.

Note: Christian Popow refers to guideline representation in general in which he has been investigating for years.

Protocure Expert Questionnaire (Prof. Dr. M. van de Bor, pediatrician)

Answers are of two types: (1) answers in the role of a **pediatrician (clinician)**; (2) answers in the role of a **guideline developer**.

1. Are you familiar with the guideline on jaundice?

Yes No

Explanation: *I have been involved in studies investigating the consequences of hyperbilirubinemia in premature newborns, and I have been in contact with the jaundice guideline developers of the American Academy of Pediatrics. Furthermore, I have been involved in the development of a guideline of hyperbilirubinemia in newborn babies for the Netherlands.*

2. What do you think about the quality (presentation and content) of the guideline?

Poor Acceptable (1&2) Good Excellent

Explanation: *The scientific evidence of the guideline is not always sound. For example, studies investigating the consequence of hemolysis in newborns indicate that the dangers for kernicterus of the levels of serum bilirubin mentioned in the guideline may have been overestimated. However, in general the guideline reflects the current state of the art, and in the context of Dutch health-care there is no need to include more tests. This may not be true for the American situation, though, as in the USA there is a significant risk that newborns are not checked on a regular basis or even escape any checks. More tests may therefore be required here, as they may not be done at all otherwise.*

3. Consider the following list of uncovered flaws. Indicate for each flaw separately how serious the flaw is in your opinion.

Incompleteness: *undefined rate of rapid TSB increase* (item 1, 'Management of hyperbilirubinemia in the healthy term newborn by age' section)

Extremely serious Very serious Serious (1&2) Not serious

Explanation: *The rate of increase should have been defined, as in the early days after birth there is normal increase in bilirubin levels; to distinguish this normal rise from abnormal, more rapid rise requires knowledge of this process.*

Inconsistency: *are clinically jaundiced children of less than 24 hours healthy newborns?* (line 1, item 1, section 'Management of hyperbilirubinemia in the healthy term newborn by age', caption Table 2, and heading flowchart are inconsistent)

Extremely serious Very serious Serious (2) Not serious (1)

Explanation: *From the viewpoint of the clinician, this is not a serious problem, as the text and tables mention that the guideline should not be used for such newborns. From the point of view of a guideline developer it is a serious flaw as it indicates lack of care in designing the guideline, which in these parts has been given a structure that looks illogical.*

Inconsistency: *signs to rule-out underlying disease* (table 1, item manifestations of the disease, versus box 2 in the flowchart, and item 7 'Evaluation' section text)

Extremely serious Very serious Serious (1&2) Not serious

Explanation: *In particular the fact that the information in the flow chart is incomplete*

is troublesome, as this renders the guidelines less suitable for inexperienced doctors (they may miss cases due to the incompleteness).

Termination: for jaundice patients who are no emergency cases, observing bilirubin levels goes on for ever (it is not explicitly stated when this can stop)

Extremely serious Very serious Serious (2) Not serious (1)

Explanation: *This will not be a problem for clinicians, as it is obvious to them that there is no need to continue requesting bilirubin levels for patients. However, it is still something that should have been dealt with in the guideline.*

Intentions:

- Phototherapy intensive treatment: **treatment intentions are only satisfied when a treatment plan is actually carried out and not stopped before the final aims are achieved (this was a result of application of the formal proof methods)**

Extremely serious Very serious Serious Not serious

Explanation: *We still do not understand what to do with this as it has no relationship with the original guideline.*

- MAJIC indicator no 7: **no more than one serum bilirubin level drawn after phototherapy is discontinued** (more than one test consistent with guideline)

Extremely serious Very serious Serious (2) Not serious (1)

Explanation: *The clinician will normally only request serum bilirubin levels of the patient if there are good reasons for it, and the strict MAJIC recommendation are therefore of little value to the clinician. However, the fact that this indicator is violated by the guideline undermines its trustworthiness. Therefore, flaws like this should be avoided by guideline developers.*

Unused test results: child blood typing (ABO, Rh) and Coomb's test (box 11 flowchart, item 1-3 text, section 'Evaluation')

Extremely serious Very serious Serious Not serious

Explanation: *These tests are used in deciding whether the rest of the guideline can be applied or that other disorders should be considered. This is explicitly mentioned in the guideline. Therefore, I do not agree with the statement that the mentioned tests are unused.*

4. Would you regard the computer supported systematic analysis of flaws an improvement over current practice?

(1) Clinician: this type of decision support will have little or no consequences for clinical practice; (2) Guideline developer: computer-supported flaw detection may improve the quality of clinical guidelines, and this will be worthwhile.

5. Please mention three pros and cons of our approach.

Pros:

- 1. useful for detection of pitfalls and ambiguities*
- 2. may give rise to production of better definitions*
- 3. more straightforward and complete protocols/guidelines*

Cons:

1. *strict protocols/guidelines are violated by clinicians more easily.*
2. *strict protocols/guidelines may give rise to 'cookbook medicine'*
3. *too much detail, and too little emphasis on medically relevant issues*

6. What is your opinion about the general methodology (formalisation of protocol and verification of properties of the protocol)?

I think personally that this approach may improve clinical guidelines considerably. It is important though that the clinician should be pointed out that it is not always mandatory to follow a guideline's recommendation literally, as guidelines offer insufficient flexibility to be always applicable to individual patients.

7. How would you judge the potentials of this approach?

Very good, in particular for guideline developers.

D5 Evaluation results, technical part

Version 1.0

December 2, 2002

Contents

1	Introduction	3
2	Selection of protocols (arrow 1)	4
2.1	Description	4
2.2	Our achievements	5
2.3	Lessons learned	5
2.4	Open issues	7
3	Informal protocol (box 2)	8
3.1	Description	8
3.2	Lessons learned	9
3.3	Open issues	9
4	Modeling (arrow 3)	9
4.1	Description	9
4.2	Our achievements	9
4.3	Lessons learned	9
4.4	Open issues	10
5	Asbru Plans (box 4)	11
5.1	Description	11
5.2	Lessons learned	11
5.3	Open issues	12
6	KIV formalisation of protocol (arrow 5)	12
6.1	Description	12
6.2	Our achievements	13
6.3	Lessons learned	13
6.4	Open issues	14
7	KIV representation (box 6)	15
7.1	Description	15
7.2	Lessons learned	15
7.3	Open issues	16

8	Identification of properties (arrow 7)	16
8.1	Description	16
8.2	Our achievements	16
8.3	Lessons learned	17
8.4	Open issues	17
9	Informal protocol properties (box 8)	18
9.1	Description	18
9.2	Lessons learned	18
9.3	Open issues	18
10	Asbru properties (box 9)	18
10.1	Description	18
10.2	Lessons learned	19
10.3	Open issues	19
11	Formal Semantics (box 10)	19
11.1	Our Achievements	19
11.2	Lessons Learned	19
11.3	Open issues	20
12	Formalisation of properties (arrow 11)	20
12.1	Description	20
12.2	Our achievements	21
12.3	Lessons learned	21
12.4	Open issues	21
13	KIV properties (box 12)	21
13.1	Description	21
13.2	Lessons learned	22
13.3	Open issues	22
14	KIV verification (arrow 13)	22
14.1	Description	22
14.2	Our achievements	22
14.2.1	Verification of Properties	22
14.2.2	Improvement of Formal Semantics	23
14.2.3	Application of Proof Methodology	23
14.2.4	Tool Support	23
14.3	Lessons Learned	23
14.4	Open Issues	24
14.4.1	Verification of Properties	24
14.4.2	Methodology	24
14.4.3	Tool Support	24
15	Conclusion	25

1 Introduction

Deliverable D5 consists of two parts. The first part describes an evaluation by medical experts and the second part is an evaluation of the project by the project team. The structure of this part of deliverable D5, devoted to the technical evaluation of Protocure project, will follow the formalisation & verification process as we described it in the original project proposal. Figure 1 illustrates this process.

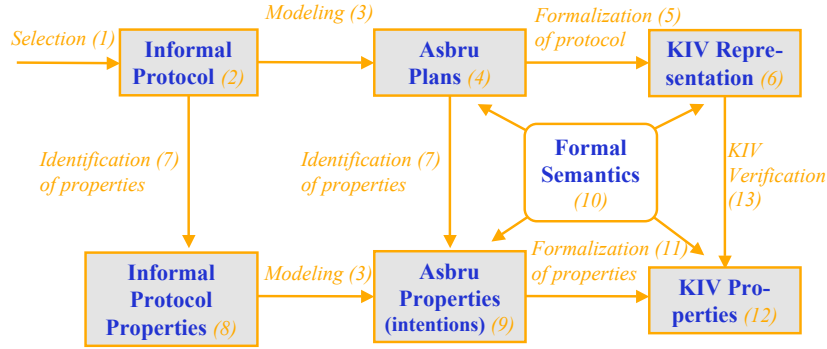


Figure 1: The whole process of formalisation and verification.

First we have selected two medical protocols which together cover a wide range of protocol characteristics (see *Selection (1)* step in figure 1). Then the two selected protocols have undergone a gradual transformation into a formal representation. Starting from the original texts (*Informal Protocol (2)* in the figure), the protocols have been first modeled in the Asbru language (*Modeling (3)* step) and then translated into the KIV formal representation (*Formalization of protocol (5)* step). The results of this transformation process are, respectively, a collection of Asbru plans representing the original protocol (see *Asbru Plans (4)*) and a set of KIV programs encoding these plans (see *KIV Representation (6)*). In order to make this formalisation possible, a formal semantics has been defined for the main constructs of the Asbru language ((see *Formal Semantics (10)*)). This is a crucial point in the process, since the formal verification we aim at is only possible with precise semantics.

In parallel to the above transformation of reference protocols, a number of interesting properties has been identified from an analysis of both the original protocols and the Asbru versions thereof (*Identification of properties (7)* step). The result of this identification phase comprises both protocol-dependent properties (see *Protocol Properties (8)*), defined at the conceptual level, and protocol-independent ones (see *Asbru*

Properties (9)), defined at the implementation level. Then we have selected a subset of properties from the previous list, and we have carried out the necessary proofs to verify them (*Verification (13)* step). Finally, the results of this formalisation & verification process have been presented for evaluation to a group of medical professionals, in order to assess the utility of our approach. The latter is the theme of the other part of deliverable D5, devoted to the evaluation of Protocure project by medical experts.

Following the process illustrated in figure 1, we will give our assessment on each of its steps (i.e. arrows), together with an evaluation of the results obtained from each of them (i.e. boxes). Afterwards, we will present a general assessment of the activities in this Protocure project. Finally, we will conclude with a summary of the techniques that we have explored or developed during this Protocure project, which we believe must be considered for a comprehensive technical evaluation.

Regarding the assessment of the different steps of the formalisation & verification process, we will present (1) a description of the activities we have carried out; (2) a concise list of our achievements; (3) a summary of the different observations we have gathered during the process (e.g. time investment, difficulties, and so on), which will be grouped under the label “lessons learned”; and, lastly, (4) a list of open issues. The latter will include aspects like e.g. short&long term improvement prospects and/or any challenging issues deserving further research. Concerning the results of each step, we will give (1) a concrete description of the (intermediate) result, (2) some comments on the lessons learned, mainly about the strong points and weak points of the result; and (3) a list of open issues.

The rest of the document roughly consists of a number of sections, one for each numbered step in figure 1, and each of them followed by a section about the results of the corresponding step.

2 Selection of protocols (arrow 1)

2.1 Description

We selected two reference protocols, which were used during the whole project to illustrate the functionality as well as the strengths and limitations of our project:

- **Jaundice protocol** 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.
Protocol (<http://www.aap.org/policy/hyperb.htm>) included in the repository of the National Guideline Clearinghouse (<http://www.guideline.gov/index.asp>).
- **Diabetes protocol** G.E.H.M. Rutten, S. Verhoeven, R.J. Heine, W.J.C. de Grauw, P.V.M. Cromme, K. Reenders, E. van Ballegooie, and T. Wiersma. NHG-Standaard Diabetes Mellitus Type 2 (eerste herziening). Huisarts en Wetenschap, 42(2):67-85, 1999.
Protocol (in Dutch: <http://nhg.artsennet.nl/standaarden/M01/start.htm>) developed by the Dutch Association of General Practitioners (NHG-Nederlands Huisartsen Genootschap).

The selection process was based on two views. The first view was language-independent based on several sources describing particular features of guidelines and protocols. The second view is based on our modeling experience and on those features we considered important.

2.2 Our achievements

The following issues have been achieved:

- We have chosen two guidelines.
- We have chosen two good quality guidelines (but one outdated).
- We have domain experts available to help us structure and translate two reference protocols into Asbru as well as KIV.

2.3 Lessons learned

We had different sources to browse or to select appropriate protocols and guidelines, which resulted in a quite huge number of possible candidates. The benefit of having a huge number of possible candidates was that we could really select out of many possibility. However, the important drawback of this huge number was, that we had to come up with selection criteria for choosing the appropriate reference protocols, which sounds simple, but was quite a difficult tasks due to complexity of guideline representations, structure of databases, time limitations, etc. (see lists of problems below). Therefore, we need a better or more appropriate list of search and selection criteria. The following list summarizes different dimensions of search and selection criteria, which limited our selection and should be changed in the future.

medical dimension:

- missing quality of the protocol
- outdated protocol (more than 5 years old)
- not evidence-based

technical dimension: the particular features of the guideline- or plan-representation language Asbru are a very skillful starting point, however, we should try to focus - also in the technical dimensions - on a more language-independent approach. For example, the Asbru- oriented features were:

- hierarchical decomposition
- knowledge rich control structure:
 - compulsory and optional plans
 - temporal order: plan layout
 - conditions
 - intentions
 - preferences and resources
 - effects
- temporal dimensions
- temporal abstraction, context

structure of a guideline: more structured guidelines needed: in the sense of algorithms, flow chart, evidence tables.

domain experts available:

- Easy to select guidelines, where you have experts (friends) in our surrounding
- Never choose a guideline without a medical person (best a guideline expert)

It is very important to use appropriate sources, databases, or libraries to search for the reference protocols. Therefore, the following list is a good example of excellent sources, recommended by the guideline and protocol developers.

Search engines:

- Google: <http://www.google.com/>
- Metacrawler: http://www.metacrawler.com/index_power.html
- Copernic(download): <http://www.copernic.com/products/copernic/index.html>

Databases:

- National Guideline Clearinghouse: <http://www.guideline.gov/index.asp>
- Tripdatabase: <http://www.tripdatabase.com/>
- University of Texas advanced searching: SUMSEARCH: <http://SUMSearch.UTHSCSA.edu/searchform4.htm>
- e-guidelines: <http://www.eguidelines.co.uk/>
- Omni-database: <http://omni.ac.uk/>

Guideline-developing organizations (the most important)

Denmark :

Danish College of General Practitioners <http://www.dsam.dk>

Finland :

Finnish Medical Society Duodecim <http://www.duodecim.fi>

France :

Agence Nationale d'Accréditation et d'Évaluation en Santé (ANAES) (before 1997 ANDEM) <http://www.anaes.fr>

French Federation of Comprehensive Cancer Centres (FNCLCC) <http://www.fncfcc.fr>

Germany :

Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften (AWMF) <http://www.awmf.de>

Italy :

Agency for Regional Health Services (ARHS) <http://www.assr.it>

The Netherlands :

Dutch Institute for Healthcare Improvement (CBO) <http://www.cbo.nl>

Dutch College of General Practitioners <http://www.artsen.net>

Sweden :

Swedish Council on Technology Assessment in Health Care (SBU) <http://www.sbu.se>

Switzerland :

Swiss Medical Association <http://www.fmw.ch>

United Kingdom :

Centre for Health Services Research Unit of University of Newcastle upon Tyne (North of England) <http://www.ncl.ac.uk/chsr>

Royal College of Physicians of London <http://www.rcplondon.ac.uk>

National Institute of Clinical Excellence (NICE) <http://www.nice.org.uk/>

Scottish Intercollegiate Network (SIGN) <http://www.sign.ac.uk>

Australia :

National Health and Medical Research Council (NHMRC) <http://www.health.gov.au/hfs/nhmrc>

Canada :

Canadian Medical Association <http://mdm.ca/cpgsnew/cpgs/index.asp>

Cancer Care Ontario Practice Guidelines Initiative <http://www.cancercare.on.ca/ccoppi>

New Zealand :

New Zealand Guidelines Group <http://www.nzgg.org.nz>

USA :

Agency in Healthcare Research and Quality (AHRQ) (before 1999 AHCPR) <http://www.ahrq.gov>

US Preventive Service Task Force <http://www.ahrq.gov/clinic/uspstfix.htm>

National Institutes of Health (NIH) <http://consensus.nih.gov>

National Comprehensive Cancer Network <http://www.nccn.org/guidelines.htm>

Institute for Clinical Systems Improvement ICSI <http://www.icsi.org/guidelst.htm#guidelines>

2.4 Open issues

(Short/long term improvements)

Define Search and Selection Criteria according – (medical oriented):

- evidence-based vs. consensus-based Guideline
- language the protocol written in
- date of publication
- databases, where looking (website, Clearing House (disease oriented))
- multiple vs. single-source
- indicators
- multi-disciplinary vs. mono-disciplinary
- support the process of diagnosis and treatment

- target groups (GP, specialists, nurses, paramedic)
- type: guideline, standards, protocols
- main objective stated
- patient's perspectives are mentioned in guideline

Define Search and Selection Criteria according – (technical oriented):

- criteria covering the whole monitoring and planning/process modeling tasks
- digital/electronically available

Challenging aspects

- How to communicate the knowledge from the technical experts to the medical experts (and the other way around).
- Providing tools for a better communication between the two experts' groups.
- Ways to combine and compare different sources of guidelines (e.g., for finding properties).
- Develop criteria for protocols that would have benefit from our techniques.

3 Informal protocol (box 2)

3.1 Description

Jaundice protocol 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.

Protocol (<http://www.aap.org/policy/hyperb.htm>) included in the repository of the National Guideline Clearinghouse (<http://www.guideline.gov/index.asp>).

Characteristics of protocol:

- not evidence-based
- algorithm
- indicators
- outdated
- text, tables, diagrams

Diabetes protocol G.E.H.M. Rutten, S. Verhoeven, R.J. Heine, W.J.C. de Grauw, P.V.M. Cromme, K. Reenders, E. van Ballegooie, and T. Wiersma. NHG-Standaard Diabetes Mellitus Type 2 (eerste herziening). *Huisarts en Wetenschap*, 42(2):67-85, 1999.

Protocol (in Dutch: <http://nhg.artsennet.nl/standaarden/M01/start.htm>) developed by the Dutch Association of General Practitioners (NHG-Nederlands Huisartsen Genootschap).

Characteristics of protocol:

- not explicit evidence-based
- no algorithm
- mainly text and tables

3.2 Lessons learned

Above mentioned guidelines were not explicit evidence-based and of medium quality according the AGREE-instrument for assessing quality of guidelines. Therefore, we need a better or more appropriate list of search and selection criteria to select guidelines for this kind of projects. Next to this it is very important to use appropriate sources, databases, or libraries to search for the reference protocols. Therefore, see list, mentioned above.

Strong points of the selected protocols are: they are from different organisations (American Academy of Pediatrics and the Dutch Association of General Practitioners), they have very different characteristics, and very important we had access to an expert for each protocol.

3.3 Open issues

In the future find a structured way to select recently published, evidence-based (EBM) guidelines or protocols of good quality and even better use semi-formal and/or formal methods during guideline development instead using it on already finished guidelines.

4 Modeling (arrow 3)

4.1 Description

The process of constructing an Asbru representation from an original medical protocol.

4.2 Our achievements

First of all we have constructed two detailed Asbru models of complete protocols. To our knowledge there are no complete protocols formalised in a protocol language like Asbru. During the Asbru formalisation of a protocol, anomalies have been discovered [11], and a categorisation of anomalies has been constructed. Beside modeling the protocols in Asbru, even implicit intentions of these protocols have been explicitly formulated in Asbru. This of course due to collaboration with our domain experts.

For the Asbru formalisation a number of utilities were developed: an intermediate Asbru language that enables us to give an Asbru formalisation, an XML-version of Asbru that enables us to build an interpreter for Asbru protocols, an interpreter for Asbru Light has been implemented, several pretty printers for several levels of detail of Asbru models has been implemented. Further more the introductory material about Asbru has been improved [12].

4.3 Lessons learned

Below we discuss the lessons that we have learned by modeling two protocols in Asbru.

- Modeling a protocol in Asbru is very time intensive (3-6 person months per protocol).
- The Asbru language is a rather complex language. Learning Asbru is difficult, and current resources are not sufficient. During the project the introductory material about Asbru has been improved.

- The required time for modeling a protocol in Asbru decreases with practice and examples. The most recent protocol outside project is done in 2 person months of student time. This was mainly due to the already available Jaundice and Diabetes Asbru models.
- modeling Asbru gets better with improved documentation and clarified semantics. Since Asbru is a very powerful language, its syntax is very complex. The documentation of Asbru consists of (1) papers in various publications describing selected aspects of Asbru (2) a manual of about 170 pages, (3) an online reference of the language elements (on the XML-level). Although the reference manual is elaborate, a full reference example and a step-by-step introduction into Asbru starting from the very beginning and gradually proceeding to the complex details is missing. This should also take into account problem-specific examples on various levels of experience of the users.
- For modeling a protocol one has to make modeling choices. This is hard, because of the lack of modeling guidelines/method.
- It was impossible to create an Asbru model from the original protocol in one step. The step between the original protocol (text, tables, free style flow charts) and Asbru language is too big for a single step. A more gradual modeling approach is needed, in other words we need intermediate models.
- Operational semantics (interpreter) helps for understanding the model during building the model.
- During modeling it is very important to have a good overview of the Asbru model. We have introduced graphical notations (e.g. tree-decomposition of plans), which were necessary. This notation and the pretty printers that we developed were necessary for realising Asbru models.
- Some medical background knowledge is required for making an Asbru plan. Besides learning the Asbru language, the modeller has to understand the protocol in medical sense.
- Asbru models are hard to understand for medics.

4.4 Open issues

Multi-sources & background knowledge A difficulty arises in places where the information obtained from a domain expert appears in the Asbru formalisation of the guideline. In theory, we wanted the guideline "as it is". In practice, we had to ask experts in several cases how to interpret details or to solve open issues. This is not a very serious problem for "normal" users of guidelines, since guidelines are designed for a medically educated audience, but it breaks the assumption that a single guideline formalized in Asbru originates from a single source. There can be several sources, just like a physician uses additional sources of information besides the guideline, too.

However, there is no efficient way to annotate pieces of the guideline in the Asbru version used in the project, yet. So some extension of the language syntax and some additional editing tools which ease the merging of information from different sources while preserving information about the source for each piece of information in the guideline are needed.

The described problems also appear in practical guideline application, when an institution like a hospital adapts a guideline to the local situation, thereby merging the original guideline with recommendations and rules from different sources.

Living guidelines Another problem is how to react to changes in the medical knowledge. Such changes occur much faster than the usual guideline formalizing process. So computer support for change over time is needed in addition to support for multiple parallel sources.

Model parts of the protocol Another issue is whether we should really model the whole protocol, or only critical (or suspected) parts. For instance in the Diabetes protocol is the treatment part the most interesting part.

Asbru construct "effects" The Asbru construct "effects" which models the expected effects of a plan are outside Asbru light. However effects are candidates for properties that should hold for a protocol. These effects have to make explicit for the protocol, in a similar way as what has already been done for intentions.

Methods/tools for supporting the modeling process Providing concrete guidelines for modeling and more tool support will reduce the modeling time. We can think about several tools, for instance visualisation tools, tools that make the modeling decisions explicit (like Guideline Markup Tool), tools that support distributed modeling and versioning of protocols. A Method for supporting the modeling are the development of design patterns. The question is here how to express recurring medical patterns in Asbru. Another way of supporting the process is construction of suitable intermediate representations and/or graphical representations. There is also a need for methods of presenting Asbru models to medical specialists.

5 Asbru Plans (box 4)

5.1 Description

A representation of a medical protocol in terms of a hierarchy of plans, including a control structure on when to execute each plan. Each plan is annotated with applicability and interrupt conditions (which can be time dependent), and with intentions describing the intended effects of executing the plan.

5.2 Lessons learned

Asbru differs on four important aspects with other medical protocol languages: (1) the hierarchical and component-based approach, (2) the use of time-annotation, (3) modelling intentions (4) the precision of the language. The hierarchical approach corresponds with the medical protocols that we have used. We can exploit this correspondence between the textual medical protocol and the Asbru model in the interaction with the expert. Concerning the second point, we agree that the time aspect is necessary in describing medical protocols. The expressivity of the time annotations seems sufficient, and not too complex for using them. For verification purposes is the modelling of intention very useful. The last point, the precision of the language. It turns out that

the language is formal enough for identifying some anomalies just by describing the protocols in this languages.

Beside these strong points of Asbru, there are also some shortcomings. For instance, in medical reasoning is uncertainty an important aspects. However there is no facilities for uncertainty reasoning. Another shortcoming is that it is not possible to model negative knowledge. Asbru is focussed on "what to do" and not on "what not to do". A last weak point is that there is no support for background information.

Two lessons that has to do with the modeling results are: the explosion in size of the Asbru model compare to the original protocol, and the definition of Asbru Light turns out to be sufficient for modelling the Jaundice and Diabetes protocols.

5.3 Open issues

The main open issue is the question "where does the explosion in size of the Asbru model comes from?" One way to analyse this increase of size/complexity is to make the modelling choices explicit. This means linking the parts from the original protocol to parts of the Asbru model, and analyse these links.

6 KIV formalisation of protocol (arrow 5)

6.1 Description

Once the two reference protocols have been modelled in the Asbru language, they have undergone a further formalization step upto the level of a formal representation—the logic formalism of the KIV interactive verifier. In addition to the usual benefits of making protocols more precise using a semi-formal language like Asbru, the main advantage of this final transformation into KIV is the possibility of a systematic analysis of protocols using formal proofs.

An Asbru protocol can be seen as a concurrent system, since the different plans which it is made of (representing clinical actions) may take place simultaneously, i.e. in parallel. Consequently, the most direct way to obtain KIV proof support for Asbru is to translate Asbru plans into parallel programs. Thus, this second (and final) step of protocol formalization has been approached as a mapping of Asbru plans to KIV programs.

However, some Asbru operators are different in nature from usual parallel programs. As result, the translation of Asbru plans to pure parallel programs can be sometimes very complex and require some encoding of the control flow of Asbru operators. This could have a negative impact on the formal verification we aim at. First, constructing formal proofs from programs with complex encodings would be harder. Second, and maybe more important, the conceptual gap between the Asbru protocol and the KIV version thereof would be too large, making difficult to trace errors discovered in the latter back to the former. In order to avoid these problems, some additional operators for non-standard parallel programs have been implemented in KIV. With the help of such operators, it is possible to translate most of Asbru features in a more comfortable way, keeping the gap between the Asbru protocol and its KIV version reasonably small.

For more information on the above topics [6].

6.2 Our achievements

The concrete achievements of this phase of KIV formalization of protocols are summarised in the following paragraphs. As main result, we have formalized the Jaundice and the Diabetes protocols in terms of the KIV representation formalism. The actual KIV formalization can be found in the appendix part of deliverable D3'. As it has been referred before, a number of additional constructs for a more direct translation of Asbru plans have been implemented in KIV. The KIV translations obtained in this way, although not perfect, can be formally verified and, at the same time, resemble their Asbru counterparts, keeping the conceptual gap relatively small. This is an important result of the formalization phase, since it means that the Asbru to KIV translation will be easier in the future, and that it could be improved further with the same strategy. Therefore, we have demonstrated that it is possible to translate Asbru protocols into KIV, obtaining formalizations of good quality. This is also an important result, since it constitutes one of the hypothesis of the Protocure project.

Lastly, the formalization process at some points contributed to/needed of closely related aspects. These include both the identification of preliminary patterns for the translation of Asbru plans and the improvement of initial Asbru formal semantics. Concerning the former, we have identified a number of Asbru schemas which appear often throughout our two reference protocols, and devised a set of standard translations (the so called translation patterns) for them, to be used as templates in similar situations. Besides, we have implemented the first version of an Asbru-to-KIV translator which makes use of these patterns. In what the Asbru formal semantics regards, the formalization of the two protocols has served to refine the version we started with. Actually, the translation of parts with which we had problems usually pointed to Asbru elements with unclear semantics, which then were clarified.

6.3 Lessons learned

The formalization of Jaundice and Diabetes protocols has served us to gain more insight on the process. We can say that now we have much more clear ideas about how to translate Asbru protocols into KIV, e.g. the (implicit) control structure of Asbru plans, the patient data, etc. In addition, we can highlight the following lessons learned:

The KIV formalization is close to the Asbru model Although the Asbru-to-KIV translation is not straightforward, the gap between the Asbru protocol and the corresponding KIV version is narrow compared with the conceptual distance between the informal protocol and its Asbru version. In addition, the difficulties of the Asbru-to-KIV translation are technical in nature, in the sense that as more technical issues will be solved the translation to KIV will become easier, more amenable to automation and thus faster. However, the modelling of the informal protocol in Asbru is inherently hard.

Formalization as (manual) translation can be improved The manual translation we carried out initially is problematic, because it is a labour-intensive and error-prone task. However, the analysis that this kind of activity requires helps identifying general translation patterns to apply to certain Asbru schemas. Thus, the formalization of Jaundice protocol gave rise to a number of translation patterns that were applied subsequently in the transformation of Diabetes. As it has been mentioned before, we

implemented an Asbru-to-KIV translator using these patterns. Actually, the first version of Diabetes formalization was mostly automated. Although this first formalization contained errors (derived from problems with our preliminary patterns), the process involved considerably less effort. We are confident that, with a proper analysis stage to improve these patterns (and define additional ones), the Asbru-to-KIV translation could be automated to a high degree. An appropriate analysis should include as well the necessary proofs of the correctness of translation patterns.

Bridging the gap between KIV and Asbru requires KIV extensions As it has been mentioned, the plan-to-program mapping was not straightforward: a number of Asbru features could not be directly mapped to parallel programs and therefore needed additional (and verbose) encoding. In order to reduce the resulting overhead in proofs, KIV was extended with a small number of Asbru-oriented constructs. This was a labour-intensive task which was not initially planned within this Protocure project. Nevertheless, it significantly reduced the effort necessary for proofs. As it is indicated in deliverable D3', in proof trials with some "toy" Asbru plans, the size of proofs without using the specific Asbru constructs were several orders of magnitude larger than the proofs using them. Our claim is that more direct support of Asbru in KIV, although requires significant effort, is essential for comfortable proofs.

Formalization serves to clarify Asbru semantics Finally, an important lesson is that formalization of realistic protocols helps refining the formal semantics of Asbru. Whenever we encountered Asbru elements for which the semantics was unclear, we made the necessary improvements to it.

6.4 Open issues

As we have mentioned in the previous section, we have gained a very valuable insight on how to translate Asbru protocols. However, there is plenty of room for improvements, at various levels and with different degrees of complexity. Next we will shortly enumerate some possible topics for future work, making the distinction between short&long term improvements, about which we have more or less clear hints, and challenging aspects, which would require more research work.

short&long term improvements Here we can think of two different lines of action, namely solving some "easy" technical issues to facilitate the translation process, and improving the current translation mechanism. More concretely, we can cite:

- integration of patient record
- integration of (plan) control variables
- improving and extending our translation patterns
- implementing an automatic translator

challenging aspects Here we can mention very different topics:

- integration of background knowledge, e.g. patient model
- integration of time with an implicit (KIV) management
- new operators for a more direct support for Asbru in KIV
- future vision: separation of Asbru plans from the behaviour of Asbru executor, e.g.
integration of (plan) control variables with an implicit (KIV) management

- “science-fiction” vision: direct support for Asbru in KIV e.g. direct transcription and verification of Asbru plans

7 KIV representation (box 6)

7.1 Description

The result of the KIV formalization step is a KIV representation of the two reference protocols that we have used throughout this project. This KIV representation not only includes a collection of parallel programs representing the different Asbru plans, but also the data specifications necessary for the domain knowledge descriptions in Asbru. As for the parallel programs, it is important to mention that the KIV representation roughly mirrors the hierarchical structure of the Asbru protocol used as starting point, due to the plan-to-program mapping strategy we followed in the formalization.

7.2 Lessons learned

The following paragraphs describe different observations related to the nature of the KIV representations we obtained, including both advantages (the so called strong points) and limitations (weak points) thereof.

Strong points First, one of the most positive features of the KIV representation, namely that it makes explicit the difficult details of Asbru plans. Actually, Asbru is a very complex language, with many different constructs that can interact with each other in multiple ways. The effort in translating a protocol into a set of parallel programs has as a side-effect an explicit formulation of the subtleties that could go unnoticed in the Asbru model. In addition to this, the KIV programs resemble the original Asbru plans in most cases, and the overall structure of the Asbru protocol is kept in the KIV formalization, thanks to the plan-to-program mapping it has been built with. Second, the KIV representation includes precise descriptions of some aspects that are less clear in the Asbru model. For instance, the KIV protocols include a specification for the patient record, together with a set of axioms describing how certain clinical actions influence them. Although this can somehow be described in Asbru, the fact is that the kind of KIV specification we use is in itself a module, resulting in a more clear description. The same occurs with the specifications of the different data abstractions. Finally, we see as a positive point the fact that the complexity of Asbru protocols can be reduced by using a smaller number of constructs for parallel programs, for which the semantics is well understood.

Weak points The main drawback of the Asbru representation is that the gap between KIV programs and Asbru plans, although reasonably small, still exists. Another important drawback, inherent to our approach, is that we should aim at verifying Asbru protocols rather than their KIV representation. However, we are aware of the fact that this would only be possible if we undertook a long-term project with the goal of providing direct KIV support for Asbru.

7.3 Open issues

The open issues we can think of currently have to do with the above drawbacks, and have been already introduced as open issues related to the formalization process. As long-term improvement, we would like to solve the necessary technical issues to make the conceptual distance between KIV programs and Asbru protocols smaller. Finally, the vision we have in mind, consisting in verifying Asbru protocols directly in KIV, would require a long-term research project.

8 Identification of properties (arrow 7)

8.1 Description

As a prerequisite for formal verification, we had to identify properties in both the original protocol and the Asbru protocol. We distinguish properties at the conceptual level and at the implementation level. Properties at the conceptual level are independent of the language used to represent protocols, but highly oriented towards the medical content of the protocol. Properties at the implementation level are protocol-representation language dependent, but independent of the particular clinical/medical protocols themselves.

Most of the properties were obtained by analysis of the informal guidelines but some were also found looking at the Asbru version of the guideline. In addition, we examined indicators and critical points of care in guidelines.

8.2 Our achievements

We identified 7 categories of protocol-dependent properties:

1. Correctness properties. They aim at ensuring that the protocol is free from faults.
2. Safety properties. Their goal is ensuring that hazardous situations will not occur.
3. Properties about (details of) actions. Here we can refer to different aspects of clinical actions, such as the duration or doses of a treatment, or even to interactions between actions.
4. Properties about traces of actions. Sometimes it might be desirable to express properties about actions in terms of their occurrence in time, e.g. sequences of required/forbidden actions.
5. Properties about particular groups of patients. They reflect the additional properties that might be necessary to manage patients with special characteristics.
6. Properties about interactions with the management of related diseases. Protocols for chronic diseases are characterised by the existence of interactions with the management of other related conditions, which might need additional attention.
7. Properties about indicators. In some cases protocols could be checked using some sort of indicator of the expected outcomes.

On the implementation level, we categorized properties dealing with coherence and correctness as follows.

1. Termination properties. These are properties, which specify a particular point to terminate the execution of each protocol. In other words, infinite loops in the protocol must not exist, because they refer to an error in the protocol.
2. Plan body implies intentions. The intentions of a plan express the purpose of the execution of this plan in a more abstract form. The conditions and the plan body define, which steps are taken to meet these intentions. In a correct plan, the plan body and the conditions imply the intentions.
3. Redundancies of plans. Finding redundancies in subplans or conditions which never trigger would be an interesting task for the future.
4. Correct replacement of subplans. Given an existing plan library, one wants to prove that certain changes did not change the overall behaviour (or certain parts of it). This plays an important role in the adaptation of global clinical protocols for the needs of a particular site, e.g., a hospital.
5. Internal Coherence. Redundancies to examine include subplans which are never used during execution, conditions which never trigger, and time annotations which do not define a constraint

The properties identified in the original protocols can be used as a basis for intentions in the Asbru plans.

A complete list of all properties identified informally and in the protocols represented in Asbru can be found in deliverable D3 and D4.

8.3 Lessons learned

Medical people are not familiar with the concept of properties. So it is not straightforward to obtain properties of protocols from them.

A weak point of properties is that this concept does not exist in the medical world, so there is no direct mapping between the thought of physician and our formal notions. This raises new problems such as assuring the appropriateness of the properties we abstracted from the information given by the physicians.

For future guidelines it would be an improvement to define informal properties based on the original guideline and then use them to define Asbru intentions. This presumes that the identification process is performed before the modelling of the guideline in Asbru.

Although the term properties is not used in the medical world, they develop indicators. We can use them as source for formulating properties. These elsewhere developed indicators are a good source for properties that may improve the protocol by using our techniques.

8.4 Open issues

Communication of the concept of properties to physicians remains difficult. To cope with this problem, there are several potential solutions. It is unrealistic to expect physicians to have formal or computer science background. Therefore, the knowledge engineers must come towards the physicians. They can do this in various ways

- by systematic analysis of the decision points in the guideline and focussing the discussion on these

- by using cases (i.e., patient cases typical for a certain path through the guideline) to obtain the desired properties from the physician this way
- interactive guideline visualization tools. This remains as a challenge for the future.

Another open issue is the formalizing and verification of more properties which are already identified. (E.g., time oriented properties and more indicators).

Yet another open issue is the further investigation into Asbru specific properties. This would require additional reasoning module possible implemented in KIV or some other platform.

In this project, we only identified static properties. We did not deal with dynamic properties such as those shown by sequences of actions or traces of protocol execution.

9 Informal protocol properties (box 8)

9.1 Description

The result of the identifying process is a set of properties (see section 8). The protocol has to satisfied these properties. One advantage is the clear statement of the properties. This may unveil hidden relations between different pieces of knowledge. Secondly, it is the basis of further verification.

9.2 Lessons learned

A strong point of the informal notation of properties was the possibility to stay close to the way the domain experts expressed themselves. Therefore, they were able to give feedback on the properties collected, which would have been impossible with other formalisations.

A weak point of the informal notation was that it was still a long way to obtain formal properties. Only then it became visible, which of the properties were trivially satisfied by the protocol or could not be proved with the supplied information and which were suitable for our methods.

9.3 Open issues

Although it contradicts the term informal properties, some framework to both guide the later transformation process to formal notation and to categorise the collected properties could prove useful if developed with care according to the practical needs.

10 Asbru properties (box 9)

10.1 Description

We produced a set of properties, which were described in an Asbru-specific way. They were expressed as intentions of plans. An intention describes either a value of a certain parameter or an action by means of a plan name. Both can either be achieved by execution of the plan, or maintained or avoided during a specified time period.

Asbru is the only protocol modelling language, which allows the user to explicitly state the intentions of plans. Therefore, the properties, which are expected to hold during the execution of a plan, can be defined directly in the plan library.

10.2 Lessons learned

From the list of potential candidates, only a part was modelled. The other was excluded because it involved background knowledge not given in the guideline. To allow for such properties to be modelled, an additional domain knowledge base or more comprehensive guidelines would be needed.

Many properties were either trivially satisfied or trivially unsatisfied by the protocol. This is not a flaw in formal verification, which precisely aims at the verification of non-trivial properties.

Asbru enable us to model one particular type of the properties that we have identified, namely the intention of a plan. In the identification phase of informal properties we identified properties which are not intentions of plans. These properties have to be modelled directly in KIV.

10.3 Open issues

The mapping of properties to KIV predicates had to be done manually. This might be changed in the future, at least for frequent forms of properties.

11 Formal Semantics (box 10)

11.1 Our Achievements

We have defined a formal semantics for the core concepts of Asbru. The semantics is operational and is notated in a Structural Operational Semantics (SOS) style. The Asbru elements which were defined mostly cover the elements occurring in the Asbru models of the selected case studies.

The SOS rules are very good for designing proof rules, and well suited to verify Asbru plans.

However, the SOS style notation turned out to be quite difficult to understand for people outside of the formal methods community, a more intuitive overview of the semantics is given using a graphical statecharts notation. Although (or maybe because) this graphical representation doesn't cover all the details, it is very good for further discussion of the intended Asbru model behaviour. It facilitates communication of the semantics to medical people without formal methods background and serves well as a language documentation for Asbru.

11.2 Lessons Learned

Asbru is a rich language containing a variety of concepts such as a hierarchy of plans, parallelism, time. Furthermore, Asbru supports e.g. cyclical plans, intentions, and time annotations. Other features are propagation, and data abstraction for monitoring measured data over time. These concepts are motivated by practical issues.

A first step was to sort out core concepts of Asbru which are relevant for the Jaundice and Diabetes case studies. This led to the definition of Asbru Light.

In order to pin down the detailed semantics of every concept in Asbru Light discussion about the precise meaning of each concept was required. This involved all project partners, but especially the Vienna group, which are the developers of Asbru. During this process it turned out that our way of notating Asbru semantics using SOS-rules was not very well suited for discussion with our partners - they simply lacked the formal background. So we had to think of another, more intuitive way of representing the semantics, which we found in statecharts. Just like Asbru has to provide a way to doctors and medically trained people to formulate treatment plans in a way they understand, we had to (re)formulate the semantics in order to make it more easily understandable.

11.3 Open issues

As a short term improvement, the formal semantics of Asbru could be completed to also cover less frequent concepts, e.g.

- local variables and return values,
- retrial of aborted plans,
- complex "wait-for" constructs, and
- iterative plan execution.

Also the behaviour of the data abstraction unit should be formally described.

In the long run further case studies should help to validate and improve the Asbru semantics. Asbru is a rich and complex language. As a consequence its semantics is complex as well and needs to be validated. The semantics will become more structured and clear and remaining errors will be corrected.

Asbru has many ways of expressing the very same thing. While this is helpful for writing concise Asbru plans, it complicates and lengthens the semantics of Asbru very much. Asbru was developed with practical applications in mind. It would be a challenging task to now improve the language from a theoretical point of view. If – like many programming languages – Asbru consisted of a smaller number of constructs each implementing an orthogonal concept and each easily understood, we could rewrite the now existing constructs as macros in those new constructs. Doing so we would maintain the current expressiveness and conciseness of the Asbru language, while greatly simplifying the core and thus its semantics. A positive side-effect would be both an easier validation of the semantics, and a not so steep learning curve for the users of Asbru.

12 Formalisation of properties (arrow 11)

12.1 Description

The informal properties of the protocol can be divided in properties that can be described in Asbru by the intentions, and properties that can not be described in Asbru. The formalisation of properties is the process that translates the Asbru intentions into interval temporal logic (ITL) formulae, and that models the other informal properties in ITL formulae.

12.2 Our achievements

We have translated a number of intentions and an indicator for Jaundice into a formula of Interval Temporal Logic.

We have derived translation pattern for intentions, because they are translated always the same (see deliverable D4).

While translating the property you have to think about the informal property in more detail, and this improves the informal property.

12.3 Lessons learned

We start out with properties. It was difficult to come up with right formulation of properties in temporal logic. Until now we were not able to formulate properties about actions. This is not impossible in temporal logic, but more difficult.

Formulating the properties is in general a hard task. Of course one has to understand the logic, but one have to be trained in temporal logic as well. So experience is important.

Not all properties can be formulated as intentions. This means that for additional properties the translation pattern cannot be applied. Manual translation is required for these properties.

A major difficulty is the following, in order to formulate the property, the KIV model of the protocol had to be enriched with additional variables (e.g. control variables to refer to different plan states). These additions required extra work and possibly modify the plan behaviour. In the future, these additions should be avoided.

These additional details give complicated temporal formulas. However for the intentions this is a minor problem because these translations follow the same pattern.

12.4 Open issues

The short term issues are to improve/complete patterns for translation of all the different types of intentions, and to realise an (semi)-automatic translation of intentions to temporal logic.

A challenge is to avoid additions to the KIV model of protocol for formulating a property. Formulating properties without additions to KIV model requires at least a separate patient model and a separate control of variables (see section 6) .

Although we are able to give translation patterns for intentions an ambitious plan would be to give directly KIV support for intentions. In order to avoid complex temporal formulas, intentions could be directly supported in KIV. This requires additional proof rules.

13 KIV properties (box 12)

13.1 Description

The KIV representation of the properties of the protocol (e.g. intentions). The KIV representation for a property is an interval temporal logic formula (ITL).

13.2 Lessons learned

A strong points of using ITL is that it has a standardized notation. Another strong point is that the translation of intentions is always the same. This means that the resulting formulas are very similar.

Some weak points are that the formulae contain many details, the formulae are difficult to understand and that additions to KIV model are necessary only to formulate properties.

13.3 Open issues

In the future, we would like to work on a closer correspondence between a property and the KIV formalisation.

14 KIV verification (arrow 13)

14.1 Description

In this step, the theorem prover KIV is used to formally verify, whether a given medical guideline satisfies certain properties. Properties are given as formulas in temporal logic (see Sect. 13), parallel program are used to model guidelines (see Sect. 7).

KIV is an interactive theorem prover, i.e. verification is partially automatic, but requires user interaction. The proof method is to symbolically execute the guideline and to use induction if necessary (for details see [7]).

14.2 Our achievements

14.2.1 Verification of Properties

We have considered two medical guidelines, Jaundice and Diabetes. Our medical experts have proposed 32 properties which the guidelines should adhere to. From this list of properties we selected four to further investigate with KIV. We were able to confirm two of the properties – they are definitely satisfied by the medical guideline. For the two other properties, we have found counter examples describing the circumstances under which they are not fulfilled (see [7] for details).

The properties considered for formal verification were of three different types.

Termination Does the Asbru plan terminate under any circumstances?

For Jaundice, the treatment should definitely end after a while, because the guideline is only applicable to newborns in their first weeks. This has not been taken care of properly in the Asbru model of Jaundice. The property does not hold.

For Diabetes this property is not very significant as the guideline is meant to be applied for lifetime.

Intentions Does the Asbru plan satisfy its intentions?

An Asbru plan should adhere to its intentions. We have successfully verified two intentions for Jaundice.

Indicators Does the Asbru plan satisfy additional properties?

Additional properties of interest are for example indicators. We have examined an indicator for Jaundice with the result that the guideline does not satisfy the indicator. A counter example was provided illustrating the circumstances under which the indicator does not hold. This counter example lead to three assumptions under which the indicator should nevertheless hold. The assumptions could either be included along with the indicator and the guideline to document the exceptional cases or the guideline could be improved such that the indicator is satisfied in further cases.

In total, formal verification is very strong for the systematic analysis of every possible case a medical guideline could be applied to. The verification of the selected properties also helped to validate, i.e. to find errors in the Asbru model. Finally we arrived with a more precise Asbru model of the guideline and with properties which were either confirmed or rejected. Counter examples illustrated the reason for rejecting a property.

14.2.2 Improvement of Formal Semantics

As was discovered during our verification effort, the initial formal semantics of the Asbru language was not as intended in some details. As described in Section 5.1 of [7], we needed to slightly corrected our understanding of Asbru intentions. Thus formal verification helped to improve the formal semantics of Asbru.

14.2.3 Application of Proof Methodology

Verification of medical guidelines has been the first large application for our newly developed proof method. With the proof method we apply the strategy of symbolic execution – which is well known for sequential programs – also to concurrent systems. By verifying Jaundice, we were able to evaluate and improve our strategy to approach large systems.

Symbolically executing Asbru plans always results in a similar proof structure. This observation can be exploited in the future to come up with proof patterns and to further automate proof construction.

14.2.4 Tool Support

Support for verifying Asbru in KIV was very rudimentary in the beginning. Our verification effort gave rise to a number of possible improvements, some of which we already implemented into the theorem prover. First of all support for more complex operators helped to reduce the gap between Asbru plans and their translations into parallel programs (see Sect. 2 of [6]). Secondly, with higher level proof rules and additional heuristics to apply them, the degree of automation has been improved to some extent (see Sect. 5.2 of [7]).

14.3 Lessons Learned

Formally verifying properties of medical guidelines was much harder than expected. As our method to verify concurrent systems was only recently developed, the people involved in the verification work package required training to become productive in

constructing proofs. Constructing the proofs was also more difficult, mainly because of two reasons.

1. *Insufficient tool support*

KIV is a powerful theorem prover providing a high degree of automation. However proof support for verifying medical guidelines was poor in the beginning. Efficient verification of guidelines is not possible with a generic proof engine alone. We had to spend some time to implement special support for Asbru into the tool. Further improvements still remain to be done. In theory, proofs for medical guidelines can be very automatic.

2. *Modular verification*

Our proof method allows for the construction of modular proofs. This is the prerequisite for verifying large systems such as Jaundice. However the basic mechanisms to modularize proofs are not yet powerful enough. It has been difficult to “guess” lemmas and to deal with the large number of possible cases which result from executing a concurrent system. We came up with ideas to reduce the number of cases but found no time to implement them yet into KIV (see Appendix B of [7] for a detailed report on our experiences).

14.4 Open Issues

14.4.1 Verification of Properties

Although different types of properties were examined, not all proofs were fully completed. The final proof that the Jaundice satisfies the indicator under the given assumptions is not yet finished. Also no properties were formally verified for the Diabetes guideline. Currently only properties restricting single states were examined. It would be worthwhile to examine a different class of properties for Diabetes which involves traces of actions.

14.4.2 Methodology

The investigation of *proof patterns* for classes of properties of medical guidelines should be a promising research topic. Proof patterns lead to more efficient proofs. They should be detectable at least for the verification of intentions, as different intentions are translated to similar temporal formulas (see Sect. 5.1.2 in [7]).

A challenging task would be to improve the *modular verification* of guidelines. Is there a method to come up with lemmas for sub plans which are suitable for the verification of properties for the parent plan? Can a proof for a single case be automatically generalised to other cases?

14.4.3 Tool Support

Efficient verification of medical guidelines requires good tool support. For our one year evaluation project, the current tool support was sufficient to do example proofs of interesting properties. This exercise pointed to a number of possible improvements which would contribute to efficiency. The possibilities are listed below.

The theorem prover we used during the project was KIV. As it turned out, the proof method of KIV is suitable for verifying medical guidelines in principle. Compared

to other interactive verifiers, KIV offer the most advanced proof method for concurrent systems. It would be promising though, to also try *model checking techniques* on guidelines, because some simple properties are more suitable for automatic verification. Model checking could be used for special classes of properties, while interactive verification could be applied to more complicated properties.

To immediately improve interactive verification, we could *enhance the symbolic execution* of parallel programs by refining the proof calculus. A properly refined calculus would allow for a very high degree of automation: in theory the proof strategy of symbolic execution is fully automatic except invariants and lemmas for modular proofs.

In the long run, tools could be improved as follows.

- *Support for additional Asbru operators.*

Asbru could be integrated even further by supporting additional Asbru operators directly within KIV. This would contribute to a more direct translation of Asbru to KIV and would make proofs more intuitive, but would also require an extended proof calculus.

- *Reuse of proofs.*

Very often formal proof attempts are not successful but reveal errors which require the modification of guidelines and/or properties. After modifications the affected proofs need to be repeated. Going through this cycle is necessary several times until finally all proofs are completed. For the verification of sequential programs a successful strategy is the automatic reuse of proofs after modifications. This strategy could be carried over to the verification of medical guidelines.

- *Automatic generation of counter examples.*

If a proof fails, it is often helpful to provide a counter example. For medical guidelines, we have generated counter examples manually up to now. Model checking provides counter examples automatically. KIV also supports automatic generation for algebraic data types and sequential programs. This could be carried over also to the verification of medical guidelines.

In principle interactive verification could be as comfortable as a debugger for an imperative programming language. As a vision, we could work towards building a “*symbolic debugger*” for Asbru plans. This would require at least

- to directly support Asbru syntax in the theorem prover,
- to support fully automatic symbolic execution of Asbru,
- to graphically represent the current execution state as far as possible, and
- to intuitively represent history of values to also illustrate counter examples.

15 Conclusion

This document contains the technical evaluation of the Protocure project. This concerns the whole process of formalisation & verification of medical protocols, as described in figure 1. We have shown that we have realised the whole process, and that we have found several points of interest (i.e. potential flaws) for improving the protocol. In the medical part of this deliverable “D5 evaluation of results”, it turns out that some

of these flaws are very important to detect and that medical experts have trust in the usefulness of our systematic way of analysing medical protocols.

In the project we developed several techniques for improving medical protocols.

The table below summarise the techniques that we developed and used for improving medical protocols.

Technique	Benefit	result
Asbru modeling	Anomalies, Explicit medical knowledge	[11, 9, 13, 8, 3]
Static verification	Quality of plans (static)	[5]
Simulation (Interpreter)	Quality of plans (dynamic)	[2]
Critiquing	Comparison between planned and actual steps	[10]
Property formalisation	Clarification of goals	[5, 7]
Verification	Confirmation Flaws, Assumptions, Explicit medical knowledge	[1, 6, 9, 4]

For realising these techniques several problems in terms of the process as depicted in figure 1 have to be solved. The basis of this work is the formal semantics (box 10). For the Asbru modeling there are a number of utilities used: an intermediate Asbru language that enables us to give a Asbru formalisation, an XML-version of Asbru that enables us to build a interpreter for Asbru protocols, several pretty printers for several levels of detail of Asbru models. Recently, the Guideline Markup Tool (GMT) is used for making the links between the original protocols and the Asbru models explicit.

The Protocure project is a one year evaluation project, in which our aim was to illustrate the usefulness of formal methods for improving medical protocols. This means that in some process steps we had to take pragmatic decisions. An example is the choice of our reference protocols. All the sections "open issues" show that we have real challenging questions for a follow-up project. In the table below, we give for each process just one such challenging question:

Process	Challenge
selection (arrow 1)	How to combine and compare different sources of guidelines?
modeling (arrow 3)	How to react to changes in the medical knowledge? (the so called living guidelines)
formalisation of protocol (arrow 5)	How to realise a clear separation of Asbru plans from the behaviour of the executor?
identification of properties (arrow 7)	How to deal with dynamic properties?
formalisation of properties (arrow 11)	How to avoid additions to the KIV model for formulating a property?
KIV verification (arrow 13)	Investigating of proof patterns for classes of properties of medical guidelines.

Another challenging question in a broader view is how to incorporate our formal approach in the design process of medical protocols.

References

- [1] M. Balser, C. Duelli, and W. Reif. Formal semantics of asbru-an overview. In *Proceedings of the 6th World Conference on Integrated Design and Process Technology (IDPT-2002)*, California, june 2002.
- [2] T. Bosse. An interpreter for clinical guidelines in asbru. Master's thesis, Vrije Universiteit Amsterdam, 2001.
- [3] Deliverable D1 – reference protocols and their asbru models. www.protocure.org, Feb. 2002.
- [4] Deliverable D2 – formal semantics of main asbru elements. www.protocure.org, March 2002.
- [5] Deliverable D3 – desirable/required properties of main asbru elements desirable/required properties of main asbru elements. www.protocure.org, April 2002.
- [6] Deliverable D3' – KIV formalisation of reference protocols. www.protocure.org, May 2002.
- [7] Deliverable D4 – verification of properties on the reference protocols. www.protocure.org, Oct 2002.
- [8] R. Kosara, S. Miksch, A. Seyfang, and P. Votruba. Tools for acquiring clinical guidelines in asbru. In *Proceedings of the 6th World Conference on Integrated Design and Process Technology (IDPT-2002)*, California, june 2002.
- [9] M. Marcos, M. Balser, A. ten Teije, and F. van Harmelen. From informal knowledge to formal logic: a realistic case study in medical protocols. In *Proceedings of the 13th European Workshop on Knowledge Acquisition, Modeling, and Management (EKAW-2002)*, 2002.
- [10] M. Marcos, G. Berger, F. van Harmelen, A. ten Teije, H. Roomans, and S. Miksch. Using critiquing for improving medical protocols: harder than it seems. In S. Qualini, P. Barahona, and S. Andreassen, editors, *Proceedings of the 8th European Conference on Artificial Intelligence in Medicine (AIME-01)*, number 2101 in Lecture Notes in Artificial Intelligence, pages 431–441, Dagstuhl, Germany, July 2001. Springer Verlag. ISBN 3-540-42294-3.
- [11] M. Marcos, H. Roomans, A. ten Teije, and F. van Harmelen. Improving medical protocols through formalisation: a case study. In *Proceedings of the 6th World Conference on Integrated Design and Process Technology (IDPT-2002)*, California, June 2002.
- [12] A. Seyfang, R. Kosara, and S. Miksch. Asbru's reference manual, asbru version 7.3, asgaard-tr-2002-1. Technical report, Vienna University of Technology, Institute of Software Technology and Interactive Systems, 2002.
- [13] A. Seyfang, S. Miksch, and M. Marcos. Combining diagnosis and treatment using asbru. *International Journal of Medical Informatics*, To appear, 2002.

DELIVERABLES TABLE

Project Number: *IST-2001-33049*
Project Acronym: **PROTOCOLURE**
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	2/12/2002

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.