Detecting and resolving inconsistencies between domain experts’ different perspectives on (classification) tasks

Derek Sleeman a,b,*, Laura Moss a,b,c, Andy Aiken a, Martin Hughes b, John Kinsella b, Malcolm Sim b

a Department of Computing Science, University of Aberdeen, Aberdeen AB24 3UE, UK
b Academic Unit of Anaesthesia, Pain & Critical Care Medicine, School of Medicine, University of Glasgow, Glasgow Royal Infirmary, Glasgow G31 2ER, UK
c Department of Clinical Physics, University of Glasgow, Glasgow G12 8QQ, UK

A B S T R A C T

Objectives: The work reported here focuses on developing novel techniques which enable an expert to detect inconsistencies in 2 (or more) perspectives that the expert might have on the same (classification) task. The high level task which the experts (physicians) had set themselves was to classify, on a 5-point severity scale (A–E), the hourly reports produced by an intensive care unit’s patient management system.

Method: The INSIGHT system has been developed to support domain experts exploring, and removing inconsistencies in their conceptualization of a task. We report here a study of intensive care physicians reconciling 2 perspectives on their patients. The 2 perspectives provided to INSIGHT were an annotated set of patient records where the expert had selected the appropriate category to describe that snapshot of the patient, and a set of rules which are able to classify the various time points on the same 5-point scale.

Results: Each of the 3 experts achieved a very high degree of consensus (≈97%) between his refined knowledge sources (i.e., annotated hourly patient records and the rule-set). We then had the experts produce a common rule-set and then refine their several sets of annotations against it; this again resulted in inter-expert agreements of ≈97%. The resulting rule-set can then be used in applications with considerable confidence.

Conclusion: This study has shown that under some circumstances, it is possible for domain experts to achieve a high degree of correlation between 2 perspectives of the same task. The experts agreed that the immediate feedback provided by INSIGHT was a significant contribution to this successful outcome.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

Contemporary knowledge-based systems, as their expert system predecessors [1], have 2 principal components, namely, a task-specific inference engine, and the corresponding associated domain-specific knowledge base. If the area of interest is both large and complex then it is likely that knowledge engineers will spend a great deal of time and effort producing the appropriate knowledge base (KB), and so various efforts have been made to reuse existing knowledge bases whenever possible [2]. The literature survey discusses a number of methods by which KBs can be produced from scratch including: traditional interviewing, computer-based tools which have incorporated classical psychological approaches such as card sort, systems to acquire information to support a particular problem solver, the use of machine learning in knowledge acquisition, as well as more recent attempts to infer information from datasets produced by large numbers of users of systems like Open Mind [3].

Two central problems, since the inception of expert systems, have been how to ensure that information provided by a domain expert represents the state-of-the-art, and secondly how to deal with uncertainty in such knowledge (i.e., experts may agree it is currently only possible to estimate the volume of a tumour from an X-ray to within certain tolerances). The first problem has often been addressed by choosing a gold standard, as in the EMYCIN project [1] where the results of the lab test were used; in other situations people have chosen the most effective expert available. In many situations, the choice of a gold standard is very problematic. The second problem (i.e., representing uncertainty) has been handled in a variety of ways. For example, EMYCIN [1] associated certainty factors with particular pieces of information (both facts and rules) and evolved a calculus which allows the uncertainty associated with decisions to be calculated, and then reported

* Corresponding author at: Department of Computing Science, University of Aberdeen, Aberdeen AB24 3UE, Scotland, UK. Tel.: +44 01224 272288.
E-mail address: d.sleeman@abdn.ac.uk (D. Sleeman).

doi:10.1016/j.artmed.2012.03.001
to the user. Fuzzy expert systems, which grew out of the fuzzy logic sub-field, have been a further development [4]. Subsequently, Bayesian networks have developed these ideas further, so that it is possible for decision support systems to identify a range of possible decisions and to associate each with strength of belief [5]. All these approaches provide pragmatic approaches to the handling of uncertainty associated with expertise. Clearly, however, there are different types of uncertainty associated with chunks of knowledge including the fact that even experts retain incorrect information, and further they can also misapply information. Developing techniques for capturing and refining expertise is an important sub-activity at the intersection of cognitive science and artificial intelligence.

The focus of the work reported here is an attempt to get a single domain expert to be self-critical of the knowledge which the expert has provided, and as a result of that reflection, on some occasions, to revise aspects of the information. Specifically, the work reported here requires the expert to classify hourly records for patients into one of 5 pre-defined categories; more particularly the expert is required to provide 2 perspectives on a classification task. Additionally, we have provided a system which enables the domain expert to appreciate when a particular entity has been classified differently by the 2 perspectives. Further, the tool provides the expert with support in revising one or both of the knowledge sources until a consensus is reached (or the expert abandons that particular task). This reflective process helps the expert detect some inconsistencies and hence produces a body of knowledge which contains fewer errors and inconsistencies. Thus this work addresses, in a novel way, one of the deep problems of a single expert building knowledge bases, namely the integrity and consistency of the articulated knowledge. This is an important issue as in rarefied domains it may only be possible to locate a single domain expert. We summarize below both the main features of our approach as well as other approaches for capturing expertise from single and multiple experts.

As usual we believe it is vital that this activity is grounded in a real-world task and we have chosen the classification of hourly intensive care unit (ICU) patient records; specifically the domain expert’s task was to classify records (which can contain around 60 pieces of patient information) on a 5-point A–E scale where E represents severe cardiovascular instability, and A represents a more stable situation (i.e., a patient who could be discharged from the unit).

The rest of the paper is structured as follows: Section 2 gives a survey of the cognitive science literature on expertise, and on knowledge acquisition including the important role which machine learning has played in these activities; thirdly we review cooperative knowledge acquisition and knowledge refinement systems, because they are related to the INSIGHT system which we have implemented recently. Section 3 provides a conceptual overview of the INSIGHT system which takes 2 perspectives on an expert’s classification knowledge, detects inconsistencies between them, and allows the domain expert to revise both knowledge sources to see if a consensus on the current task can be reached. Section 4 describes the use of INSIGHT by experts to reconcile 2 perspectives of their knowledge about ICU patients; namely a set of annotated patient records and a rule-set which covers each of the 5 categories (A–E). A high level of consensus was achieved between these 2 perspectives by each of the 3 experts in the study. Further we also, in a subsidiary study, achieved a high correlation between the 3 experts, by introducing a commonly agreed rule-set. Section 5 outlines several of the contributions of this work; and Section 6 concludes the paper by outlining some planned future research.

2. Literature survey

This section gives a cognitive science perspective on the acquisition of expertise (Section 2.1), provides a brief survey of knowledge acquisition (including machine learning) approaches in Section 2.2, and discusses cooperative knowledge acquisition and knowledge refinement systems in Section 2.3.

2.1. The cognitive science perspective on the acquisition of expertise

Although domain experts are highly regarded, domain experts can be susceptible to errors and biases which can affect their performance [6]. Such inconsistencies can have consequences for the quality of the information acquired during knowledge acquisition activities.

Of particular interest to this paper is the affect of context on an expert’s performance. When the context of a problem is changed (e.g., two different perspectives on the same task), or when versions of the same stimuli are rearranged, the superior performance of an expert, compared to a novice, is not always repeated [7]. For example, Lewandowsky et al. [8] found that expert bush fire commanders, when presented with very similar fires in two different contexts, made opposing predictions about the spread of the bush fire. The classic book on protocol analysis by Ericsson and Simon [9] argues that to acquire a person’s genuine expertise it is essential that one does not get the expert to articulate what they do in the abstract, but one should essentially observe what they do when solving an actual task. Effectively, Ericsson and Simon introduced the distinction between “active” knowledge which is used to solve tasks as opposed to “passive” knowledge which is used to discuss tasks or a domain. This has been a recurrent theme in much of cognitive science and in the study of expertise since that time, as is illustrated by an early study reported by Johnson [10]. This investigator attended a medical professor’s lectures on diagnosis where the professor explained the process. The investigator then accompanies the professor’s ward round (with a group of medical students) and noticed a difference in his procedures. When challenged about these differences, the professor said:

“Oh, I know that, but you see I don’t know how I do diagnosis, and yet I need to teach things to students. I create what I think of as plausible means for doing tasks and hope students will be able to convert them into effective ones.”

Thus the essential “rule” of knowledge acquisition is that one should ask an expert to solve specific task(s), and (preferably) explain what the expert is doing as the task proceeds; one should not normally ask a domain expert to discuss their expertise in the “abstract” (this includes asking an expert to articulate rules and procedures they use to solve tasks). Klein and colleagues [11] have also made important contributions to the study of naturalistic decision making; by studying experts such as firefighters in their natural environment, they discovered that laboratory models of decision making could not describe decision-making under uncertainty. Mayer [12] reminds us that expertise has many “facets” including: conceptual, procedural, and strategic knowledge. Further, Baxter et al. [13] discuss the (cognitive) task analysis which they did in an attempt to ensure that their decision support system, FLORENCE, would integrate well with the current working procedures of their neonatal ICU. (This paper would provide us with useful information, should we subsequently integrate INSIGHT into Glasgow Royal Infirmary’s ICU; currently all our analyses are being done “off-line.”)
2.2. Summary of knowledge acquisition approaches

In an overview at the K-CAP 2007 conference, Sleeman et al. [14] argued that knowledge acquisition is “a broad church” and consists of a wide range of approaches including:

- Interviewing domain experts by knowledge engineers: an approach which was dominant in the early development of expert systems [1]. (Note there are a variety of strategies which can be pursued in such interviews [15].)
- Techniques, including card sort, repertory grids, laddering, which had originally been developed by psychologists as “manual” techniques which computer scientists redeveloped as a series of computer-based systems [15].
- Problem solving method (PSM) driven systems such as MOLE, MORE, and SALT acquire more focussed information which is sufficient to satisfy a particular type of problem solver or PSM. The use of these systems is less demanding for the domain expert as the information collected is generally less extensive, and the purpose of the information collected is usually more apparent [16].
- Machine-learning approaches have played an important role of transforming sets of usually labelled instances into knowledge (usually rule-sets). An early successful application of machine learning approaches is that by Michalski and Chilausky [17]; their application was to identify diseases in the Illinois soybean crop.
- Natural language techniques (specifically information extraction approaches) have now matured to the point where they have been successfully applied to a number of textual sources and have extracted useful information [18].
- Capitalizing on greater connectivity and the willingness of some people to provide samples of texts, and to complete sentences in meaningful ways. Systems like OpenWorld have collected vast corpora which they have then analysed using statistical techniques which are able to refine KBs in a variety of formalisms including rules, cases, taxonomies, and causal graphs. The family of systems which computer scientists redeveloped as a series of computer-based systems are provided with a set of categories which the domain expert believes are relevant to the domain, a set of descriptors and hierarchical; if the latter, then the system requires some further information about the nature of the taxonomy. The medical example used in this section, is the one used initially to demonstrate the functionality of the REFINER systems. Table 1 shows a set of cases including the categories assigned by the domain expert to each case. At the heart of each system is an algorithm which

### Table 1

<table>
<thead>
<tr>
<th>Case</th>
<th>Heart rate (HR)</th>
<th>DBP</th>
<th>Disease</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50</td>
<td>90</td>
<td>Disease 1</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>56</td>
<td>90</td>
<td>Disease 2</td>
<td>A</td>
</tr>
<tr>
<td>3</td>
<td>52</td>
<td>101</td>
<td>Disease 3</td>
<td>A</td>
</tr>
<tr>
<td>4</td>
<td>50</td>
<td>95</td>
<td>Disease 3</td>
<td>B</td>
</tr>
<tr>
<td>5</td>
<td>56</td>
<td>97</td>
<td>Disease 3</td>
<td>B</td>
</tr>
<tr>
<td>6</td>
<td>89</td>
<td>89</td>
<td>Disease 5</td>
<td>A</td>
</tr>
<tr>
<td>7</td>
<td>52</td>
<td>97</td>
<td>Disease 3</td>
<td>B</td>
</tr>
</tbody>
</table>

...
forms a category description from say, all the instances of category A, bearing in mind the actual types of the variables. This process is repeated for each of the categories. Table 2 shows the category descriptions which the algorithm infers for this dataset. The systems then check to see whether the set of inferred categories are consistent (i.e., not overlapping with other categories). The set of cases is said to be consistent if each category can be distinguished from the other categories by a particular feature or a particular feature–value pair.

If the set of cases is inconsistent then the algorithm further suggests ways in which the inconsistency(s) might be removed; these include:

- Changing a value of a feature of a case.
- Reclassifying a case.
- Adding an additional descriptor to all the cases.
- Creating a disjunction by excluding a value or range of values from a category description.
- Shelving a case to work on it subsequently.

Considering the dataset shown in Table 1, the category descriptions are inconsistent (a case with a diastolic blood pressure (DBP) value in the range 95–97 and a disease value of Disease 3 could not be unambiguously categorized) and so the user would be presented with a set of disambiguation options such as:

- Exclude 95–97 from category A's DBP range.
- Change the value of disease in case 3 to Disease 1, Disease 2 or Disease 5.
- Add a new descriptor to distinguish between these categories.

If, for example, the user opts to create a disjunction, the categories are now distinct. Table 3 shows the updated (non-overlapping) categories:

We have so far effectively implemented 3 systems:

- REFINER [31] was the first system; it was incremental in that it processed a single case and attempted at each stage to remove any inconsistencies detected.
- REFINER+ and REFINER++ [32]: The clear disadvantage of REFINER was that a change made to accommodate an inconsistency associated with case \( n \) might be reversed when case \( (n + 1) \) was considered, and so REFINER+ implemented a “batch” algorithm. Namely all the instances were available before any of the category descriptions were created, and hence it was able to avoid much of the unnecessary work done in the initial system. When REFINER+ was used with a small number of cases it was quite effective, however the number of inconsistencies noted in a sizable dataset could be overwhelming for the expert. To help contain the situation we evolved several heuristics namely:
  - A modification which removes a considerable number of inconsistencies is preferred over one which removes a smaller number of inconsistencies.
  - A modification which makes a small number of changes to the dataset is preferred over one which makes a larger number of changes.
- REFINER DA: The essential difference between REFINER DA and its predecessors (REFINER+ and REFINER++) is that it combined aspects of the two earlier systems. Namely the domain expert is asked to suggest several cases which the expert thinks are prototypical of the several categories, from which descriptions of the several categories are inferred as described above. REFINER DA then attempts to cover additional cases without causing the category descriptors to become inconsistent.

REFINER’s machine learning algorithm attempts to create a set of non-overlapping descriptions for the categories; moreover, each descriptor is used in each of the categories. Further, the descriptor–value pair which effectively discriminates category A from category B is produced by the machine learning algorithm, and hence is greatly influenced by the set of cases presented to the system. The domain expert’s intuitions are not used to guide this selection of features. We report here an example which we encountered when using REFINER DA with the ICU domain. The system selected a high value for oxygen saturation as the discriminating feature for a “potentially dischargeable” patient; however if that patient is also on a high-level of inspired oxygen, then this person should be classified as a very sick patient. So from working with REFINER DA we have made two important observations:

- The feature–value pairs chosen by the REFINER systems to describe categories are often not very intuitive to a domain expert. (The same comment can of course be made of the output from other machine learning algorithms such as the decision trees created by the C4.5 algorithm [33].)
- An expert might effectively sub-divide a category like B into a number of sub-categories, which he may not initially articulate. (That is a patient can be assigned to category B for one of several distinct reasons: e.g., poor heart rate, or the existence of Disease 3.) If the domain expert does not articulate these sub-categories then category B is likely to be an amorphous category which will influence the descriptions inferred by the system for that category; this in turn will affect the other categories inferred by the REFINER systems. Further, if the sub-categories are articulated then it is likely that there will only be a small number of instances in each of the sub-categories, which again will mean that the machine learning algorithm(s) will have difficulties in extracting meaningful domain-relevant descriptions for the several categories.

In the next section we outline a further system which we have developed, called INSIGHT, which addresses these issues.

3. Conceptual design of INSIGHT

Below we give the design criteria for a system, INSIGHT, the first point below addresses the (several) difficulties noted at the end of the last section.

- Have the experts describe each of the categories and sub-categories in terms of features which the expert believes are appropriate. Effectively the expert provides a set of
classification rules for the domain (i.e., for each of the categories and sub-categories).

- All the REFINER systems require the domain expert to assign a category (a label) to each of the instances. We are continuing this practice here as it gives us a further perspective on the set of cases.
- Compare the expert’s 2 perspectives on the domain; namely, the annotations the expert has associated with each of the cases, versus the categories suggested by the rule-set when it is executed against the several patient instances.

As noted above, INSIGHT is a development of the REFINER family of systems, yet incorporates a considerably different approach.² Whereas the REFINER systems are able to infer descriptions of categories from a set of labelled instances to detect inconsistencies, and suggest how they might be resolved, the INSIGHT system highlights discrepancies in two perspectives provided by a single expert on a particular (classification) task, and brings these to the expert’s attention. Specifically, INSIGHT is able to handle annotated cases where the expert assigns each instance to one of the pre-designated categories (the first perspective) and the second perspective, a set of rules defined by the expert who classifies each of the instances.³ INSIGHT displays the results of such comparisons as a confusion matrix. Fig. 1 provides a screenshot of INSIGHT’s graphical user interface (GUI). In the left hand panel, the current rule-set is displayed (shown in greater detail in Fig. 2): and the associated confusion matrix is displayed in the right hand panel. In this example, 3 categories have been chosen by the expert: HR_extreme, HR_moderate, and HR_normal. The domain expert has annotated a set of instances using the 3 categories and has defined a set of rules (Fig. 2) to describe the same 3 categories. The comparison between the outcomes of the domain expert’s rules and the annotated data can be seen in the confusion matrix; the first row of the matrix consists of all the cases which have been annotated by the domain expert as ‘HR_extreme’ whereas the cell (HR_extreme, HR_moderate) corresponds to cases which have been annotated by the expert as ‘HR_extreme’ but have been classified by the domain expert’s rule-set as ‘HR_moderate’. Clearly all the diagonal cells [i.e., (HR_extreme, HR_extreme) (HR_moderate, HR_moderate) (HR_normal, HR_normal)] contain instances which have been classified identically by both the expert’s annotation and rule-set.

INSIGHT provides a range of facilities to enable the expert to view the instances which have been misclassified and to either edit the dataset (say to change the annotation of an instance, or correct a clearly incorrect data value) or to revise or enhance the current rule-set.

Fig. 1. A screenshot of INSIGHT’s GUI showing a set of (domain) rules and the corresponding confusion matrix. [Here ‘from data’ refers to the expert’s classification of a case, and ‘from rules’ refers to the instance’s classification by the rule-set.]

Fig. 2. The figure shows (at greater resolution) the 3 rules displayed in Fig. 1 (i.e., in INSIGHT’s GUI). This simple rule-set contains just 3 rules: the first 2 rules are disjunctive and the third has a single condition.

---

² We shall see later that one of INSIGHT’s modes does use machine learning techniques, in a limited way.

³ Approaches exist which enable the decisions of 2 experts to be compared; some of these comparisons are quite innovative, and we plan to explore whether any of these could be exploited in the INSIGHT approach. However, we know of no other system which currently enables a domain expert to compare 2 perspectives of a task directly.
Our assumption was that the confusion matrix would be a very intuitive way of presenting results to experts.⁴ So far we have observed 10 experts using INSIGHT, and from their actions we have inferred that they all understood the information inherent in the confusion matrix. Moreover, several of them commented to us informally that they thought it was a succinct way of presenting the information. Additionally, it suggests a procedure (set of heuristics⁵) for tackling revisions of the discrepancies. Clearly some discrepancies are more surprising than others. For example, as all the categories are in a sense ordered, instances in the cell \((HR_{\text{extreme}}, HR_{\text{normal}})\) can be considered to be more surprising than those only one category away, say those in cell \((HR_{\text{extreme}}, HR_{\text{moderate}})\). Thus this distance measure suggests that the domain expert should be encouraged to consider discrepancies in the following order:

- \((HR_{\text{extreme}}, HR_{\text{normal}})\) and \((HR_{\text{normal}}, HR_{\text{extreme}})\) (distance between categories of 2).
- \((HR_{\text{extreme}}, HR_{\text{moderate}}), (HR_{\text{moderate}}, HR_{\text{extreme}}), (HR_{\text{normal}}, HR_{\text{moderate}})\) and \((HR_{\text{moderate}}, HR_{\text{normal}})\) (distance between categories of 1).

The above advice was provided to the clinicians whilst they were being introduced both to the refinement task and to the INSIGHT system. We made two other suggestions as to how they might approach the task of making the two perspectives more consistent, namely:

- To initially refine each of the patient datasets individually, before attempting to refine the complete set of instances (i.e., all the patients’ datasets).
- In the first period to concentrate on removing the discrepancies from the dataset (incorrect annotations and data points) and only at a later stage make changes to the rule-set. This heuristic is based on the perspective that changes to the dataset are localized, whereas a change to a rule could, in principle, affect all the cases.

Note this advice has focused on how the expert might approach the refinement task (e.g., which aspect the expert might tackle first) and did not discuss the changes to be made, by the expert, to the annotations or to the rules.

Although we report the use of INSIGHT applied in the ICU domain (Section 4) the tool has been designed generically so that it can be applied in other domains; Section 6 describes such future work plans.

3.1. The rule interpreter

Essentially each rule consists of a single action which assigns a particular instance to a category, and a series of conjunctive conditions or a series of disjunctive conditions. (More recently we have added the ability to specify that \(M\) out of \(N\) conditions are satisfied – this has proved useful in reducing the number of rules required in several applications.) Fig. 2 provides a simple rule-set based on the ICU domain. (Section 4.1 discusses a more extensive ICU rule-set.) To date we have implemented only a single conflict resolution strategy, namely the first rule which is satisfied, fires. This means that it is necessary for the domain expert (supported by the analyst) to ensure that the most specific rules are placed at the top of the list, and the more general rules are placed lower in the list. For example, if we had the following 2 conjunctive rules: rule 1: \(A\) and \(B\) and \(C\) together with rule 2: \(A\) and \(B\), if rule 2 occurred in a rule-set before rule 1, then with the simple conflict resolution strategy we are using, rule 1 would never be activated. In many situations rules are mutually exclusive, as they include non-overlapping conditions (or in the extreme case use completely different descriptors) in which case they are order-independent. However, if a set of rules have related conditions, then it is important to ensure the rules are appropriately ordered, as the above example indicates.

We have kept the format of the rules and the rule interpreter simple for a number of reasons: firstly, this means the system could be implemented quickly; and secondly, the form of the current rules and the interpreter’s decision making appear to be easily understandable by domain experts. (The interpreter and the form of the rules may be enhanced subsequently if there is a clear need.)

3.2. Inferring rules from instances

INSIGHT has a further mode which infers a rule when provided with several instances of a particular category. This mode was added so that an expert would not be forced to specify rules for each of the categories ab initio. However, such rules contain a feature-value pair corresponding to each of the descriptors used to describe categories. Our recent work with INSIGHT has made us aware of the need for experts to select relevant descriptors from the inferred rules, in order to achieve domain-acceptable distinctions between the categories. This mode has still to be used by a domain expert with a demanding application, and as such is featured in Section 6.

4. Evaluation of the INSIGHT system

The objective of the ICU study using INSIGHT is to derive a series of rules which can be used with a high degree of consistency, to classify the hourly patient reports produced by the patient management system. The overall objective of the project is to produce a methodology and tools to support the process by which an expert attempts to make several perspectives about a task, consistent.

Many ICUs have patient management systems which collect the patients’ physiological parameters, record nursing activities, and other interventions (such as the administration of drugs and boluses of fluids). This information, typically collected at specified time periods say every minute or hour, is recorded in a database associated with the patient monitoring system, and is continuously available on a monitor at the patient’s bedside where it is usually displayed as a conventional chart; this is the form of the information which clinicians use when they attend patients. Thus many ICUs are now paperless. Often this information is not systematically analysed subsequently for trends or inconsistencies in the datasets. This is the focus of a complementary aspect of our work which led us to produce the architecture for clinical hypothesis evaluation infrastructure (ACHE) [35]. That paper also outlines one preliminary study which we have undertaken with ACHE to identify the occurrence of myocardial infarctions in this group of ICU patients.

The patient management system used at Glasgow Royal Infirmary records up to 60 parameters. Table 4 lists the principal parameters, and lists the frequency of recording in the current dataset. It should be noted that the datasets which we analyse are extracted from the patient database, de-identified, and output as a spreadsheet; the spreadsheet is then input to the ACHE system and different analyses are performed on the data “off-line”.

For a variety of reasons it is helpful for clinicians to obtain a summary of each patient’s overall condition, on request, from the data recorded by the patient management systems (currently most ICU patient management systems provide reports every hour,

---

⁴ Confusion matrices are a common method to display machine learning results [34].

⁵ The heuristics used here are not the same as those used in the previous REFINER family of systems as the tasks being addressed are different.
The main parameters used in the study.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Recorded interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>HR</td>
<td>Hourly</td>
</tr>
<tr>
<td>Temperature</td>
<td>Hourly</td>
</tr>
<tr>
<td>Mean arterial pressure (MAP or mean)</td>
<td>Hourly</td>
</tr>
<tr>
<td>Diastolic</td>
<td>Hourly</td>
</tr>
<tr>
<td>Systolic</td>
<td>Hourly</td>
</tr>
<tr>
<td>FiO₂</td>
<td>Hourly</td>
</tr>
<tr>
<td>SpO₂</td>
<td>Hourly</td>
</tr>
<tr>
<td>Urine output</td>
<td>Hourly</td>
</tr>
<tr>
<td>Central venous pressure (CVP)</td>
<td>Hourly</td>
</tr>
<tr>
<td>Drug infusions (e.g., adrenaline and noradrenaline)</td>
<td>As applicable</td>
</tr>
<tr>
<td>Fluid infusions</td>
<td>As applicable</td>
</tr>
<tr>
<td>Dialysis sessions</td>
<td>As applicable</td>
</tr>
</tbody>
</table>

but there is the possibility of more frequent reports if these are needed clinically. Such information would be useful to determine whether there has been any appreciable improvement or deterioration, would be a useful summary for the next shift of clinical staff, and could be included as a component of a discharge summary. To date the APACHE II scale [36] is widely used in ICUs in the western world; however, the APACHE score is calculated only once during a patient’s ICU stay, usually during the first 24 h after admission. Additionally this scoring system does not take into account the effect of interventions on a patient. For example, if a patient has a very low blood pressure this is clearly a very serious condition, but it is even more serious if the patient has this blood pressure despite having received a significant dose of a drug like adrenaline.⁶

The clinical authors of this paper (JK and MS) have been analysing ICU patient scoring systems for some whilst. More recently we have produced a 5-point (high-level) qualitative description of ICU patients [37], which can be summarized as follows:

- **E**: Patient is highly unstable with say a number of his physiological parameters (e.g., blood pressure and heart rate) having extreme values (either low or high).
- **D**: Patient more stable than those in category E but is likely to be receiving considerable amounts of support (e.g., fluid boluses, drugs such as adrenaline, and possibly high levels of oxygen).
- **C**: Either more stable than patients in category D or the same level of stability but on lower levels of support (e.g., fluids, drugs and inspired oxygen).
- **B**: Relatively stable (i.e., near normal physiological parameters) with low levels of support.
- **A**: Normal physiological parameters without use of drugs like adrenaline, only small amounts of fluids, and low levels of inspired oxygen.

For more details on the categories, see Appendix A.

The outline of the study to generate and refine a set of rules for the previously proposed 5-point qualitative description of ICU patients is:

(a) The administrator of the patient management system produced listings (in spreadsheet format) for 10 patients’ complete stays in the ICU (the number of days varied from 2 to 23 days). These 10 records were chosen randomly from the larger number of patient records available. (Each spreadsheet contained the parameters described in Table 4.)

(b) One of the clinical investigators (clinician-1) annotated each of the hourly records (nearly 3000 records in all) with his assessment of the patient’s status on the 5-point qualitative scale on the basis of the information provided by the patient management system i.e., that contained in the spreadsheets.

(c) Further, we asked the same clinician to articulate rules to describe each of the 5 categories, (i.e., A–E).

(d) We used the INSIGHT tool to help this clinician make this dataset and his rule-set more consistent by modifying, as he saw fit, the dataset, the dataset’s annotations, his rule-set, or any combination of these.

(e) Due to time constraints, the second clinician annotated 3 of the patients’ datasets, again using the same qualitative scale (A–E). (These datasets were randomly chosen from the 10 annotated by (clinician-1).)

(f) The INSIGHT tool was then used to help the second clinician make his dataset consistent with the rule-set produced by the first clinician. This clinician was, of course, allowed to modify the dataset, his annotations of the dataset and the rule-set.

(g) The last 2 steps were repeated with a third clinician (where the third clinician used initially the second clinician’s final rule-set).

(h) We held a session with all 3 clinicians in an attempt to agree a common rule-set.

(i) Independently, the 3 clinicians reconciled their own annotations (datasets) with the common rule-set.

In this study, highly experienced clinicians were asked to assess hourly snapshots of patients and to categorize each of the instances on a 5-point scale. Note we were not assessing the depth or organization of the experts’ knowledge nor how that knowledge had been acquired. That would be a considerably different, and a more complex, study. If we were to address this topic, particularly if we were to study how neo-experts become “real” experts in the complex domain of ICU medicine, then we would be greatly influenced by the earlier work of [38].

As noted in Section 1, INSIGHT has been developed to enable a single domain expert to make their 2 (or more) perspectives more consistent. So one should view this study as the use of INSIGHT with 3 separate experts to determine how effective it is with a number of individual experts. This is the core of the study. However, we took advantage of having access to the 3 experts to hold a face-to-face session (step h above) to see if they could reach a consensus on their several rule sets. Subsequently in step i, each individual expert used INSIGHT in an attempt to reconcile his own annotations with this common rule-set. So again the focus of the study returns to the use of INSIGHT by an individual expert.

Further, it should be noted that the patients’ datasets represent their complete stay in the ICU, and hence it is to be expected that the quality and completeness of the records will not be high at both the beginning and end of the patients’ stays. For example, usually when a patient is first admitted to an ICU, they are in need of resuscitation, and as some of this involves manual infusion of drugs, the patient management system does not capture all the activities, or all of the patient’s physiological parameters. Thus although information about the state of the patient during this period may have been communicated amongst the members of the clinical team, a considerable amount of this information will not have been captured by the ICU’s patient management system, and thus will not be available for analysis by INSIGHT. Further, associated with each patient’s stay there may be a number of time points which do not contain all the “core” parameters, and hence, it might be argued, these time points should not be used for analyses.

Note that after the first 6 h in the ICU, a complete set of “core” parameters is normally collected by the patient management system for the patient. (We have agreed, in principle, to only analyse instances after the first 6 h of a patient’s stay.) It should be noted that some of the descriptors in this dataset (such as urine output and heart rate) were extrapolated to provide certain missing values; the algorithms used to calculate these missing values were agreed with the clinicians. Further, it should be noted that the 10 cases provided for the study were randomly selected from all the
cases stored in the ICU's patient management system, and hence should represent a mixture of case severities. On the other hand, it is important to point out that the patients in this cohort come from a unit with a greater severity of illness than most intensive care units. The mean APACHE II scores of patients in this unit are in the range 19–25, whereas the national average is 19 [39].

This section describes a two-phase study conducted with clinician-1 (Section 4.1), a related 1-phase study undertaken with clinician-2 (Section 4.2), a 1-phase study with clinician-3 (Section 4.3), an initial comparison between the performances of the 3 clinicians (Section 4.4), and approaches for achieving consensus between the 3 clinicians (Section 4.5).

### 4.1. Study with clinician-1

For a number of reasons, the study with this clinician was somewhat more extensive than with the other 2 clinicians. Firstly, this was the first use of the INSIGHT system with a domain expert, and so we decided in this phase just to use a limited dataset (i.e., data from a single patient). During this first phase we realized that the INSIGHT system needed to be made easier and more efficient to use. In the second phase, clinician-1 used the improved INSIGHT with the remaining 9 patients. As part of the first phase, clinician-1 articulated an initial set of rules (which he then refined in the two phases. The study with clinician-2 took as its starting point the final rule-set produced by clinician-1).

As a result of these 2 phases with clinician-1, it was realized that the time commitment to annotate and then refine the dataset and rule-set with INSIGHT for all 10 patients was too great a time commitment for most clinicians, and hence we randomly chose a subset of 3 patients for use with the remaining clinicians. (So in a real sense the study with clinician-1 was a pilot; where we investigated the “stability” of the INSIGHT tool and the time to process and refine a single patient dataset.) When we make comparisons between the 3 clinicians we report data on the 3 common patients; additionally, in this section we report the results achieved by clinician-1 with the 10 patient datasets.

Clinician-1 chose initially to concentrate on patient 705 which has 576 time points (or instances). When he started this session there was a 45.0% (259/576) agreement between his annotations and the results obtained by running his initial rule-set against the (patient) instances, however if the unclassified instances are ignored that figure becomes 45.9% (258/562). Further, at the end of the session the agreement was 97.0% (559 of 576) or 100% (556 of 556) if we ignore the effects of the (20) unclassified instances.8 Kappa values have also been calculated for each of these calculations. Table 6 summarizes these results. Each cell gives the percentage match and the corresponding kappa value [40]; in the second row we give the number of items matched and the total number of items. We are using the interpretation of kappa values suggested by Landis and Koch [41] namely: <0, no agreement; 0.0–0.20, slight agreement; 0.21–0.40, fair agreement; 0.41–0.60, moderate agreement; 0.61–0.80, substantial agreement, and 0.81–1.0, almost perfect agreement. So the corresponding kappa values before refinement (0.24 and 0.25), with and without unclassified items, are both classified as “fair agreement”; after the refinement the corresponding kappa values are 0.95 and 1.00 and so are both classified as “almost perfect agreement”. So the refinement process has significantly altered the agreement between the corresponding dataset and rule-set.

This session took about 5 h, and was relatively slow as this was the first time INSIGHT had been used on a real application by a domain expert, and at the beginning of the session it was necessary, for example, to change annotations of instances singly, which was painstaking when the expert wanted to change a group of annotations. This and other functionalities have subsequently been added so this tool is now very much faster to use. One thing which this clinician did at an early stage was to reduce the number of parameters viewed for each instance from the original 41 to just 6: this also speeded up his handling of instances considerably. The parameters which he chose to view were: adrenaline, FiO2, HR, MAP, noradrenaline, and SpO2.

Earlier in this section, we outlined the nature of the patient data available in this domain, and in Section 3 we outlined the simple rule interpreter which we have implemented in INSIGHT. Here we give some examples of the rules which this clinician articulated for several categories. For example, the rule associated with category A has the following form:

There is insufficient information on which to base a classification or the rule-set fails to classify an instance due to lack of information. (Note the expert might also use this classification in other circumstances.) Below, we give examples of the 2 subgroups:

- An expert assigns an instance to [none] on the basis that values are not available for what he judges to be core parameters, but the rule-set assigns the instance to a category on the basis of the instance matching one of the (say extreme) disjunctive conditions associated with a domain rule. (See Fig. 2 for examples of rules for this domain.)

- An instance classified by the expert as an “A” is classified by the rule-set as [none] as several of the descriptors specified in the “A” rule do not have values associated with them. In all cases, the unclassified instance is one which has information missing which either the expert or the rule-set thinks is important. Ideally, we would introduce the concept of an “classifiable” instance, i.e., an instance which has values associated with all the core descriptors (say FiO2, SpO2, MAP, HR and all the inotrope drugs), and only classifiable instances would be further assigned to a particular (A–E) category. However, we are not able to currently handle the execution of classifiable instances as only non-zero drug doses are recorded, and the above approach would require zero (inotrope) doses also be recorded by the ICU's patient management system.

---

8 For the purposes of this study, we have defined an “unclassified” instance as an instance which either the expert assigns to category [none], as the expert believes

---

### Table 5

<table>
<thead>
<tr>
<th>Patient code</th>
<th>696</th>
<th>705</th>
<th>707</th>
<th>708</th>
<th>720</th>
<th>728</th>
<th>733</th>
<th>738</th>
<th>751</th>
<th>782</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of hourly time points (or instances)</td>
<td>129</td>
<td>576</td>
<td>475</td>
<td>40</td>
<td>188</td>
<td>281</td>
<td>396</td>
<td>110</td>
<td>493</td>
<td>73</td>
</tr>
</tbody>
</table>

---

### Table 6

Summary of clinician-1’s refinement for patient 705. Each cell gives the percentage match and the corresponding kappa value; in the second row the number of matching items and the total number of items are cited. As the rule-set was produced by the same expert, intra-rater consistency is being reported here.

<table>
<thead>
<tr>
<th></th>
<th>Patient 705 (before)</th>
<th>Patient 705 (after)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All instances considered</td>
<td>45.0%: 0.24</td>
<td>97.0%: 0.95</td>
</tr>
<tr>
<td>Unclassified instances excluded</td>
<td>45.9%: 0.25</td>
<td>100.0%: 1.0</td>
</tr>
</tbody>
</table>

---

8 Note 3 instances were classified by both the expert and the rule-set as “unclassified”. This explains the change in the number of the classified from 559 to 556 when the unclassified are removed from the calculation.

9 With 10 patients there are 41 descriptors; clinician-2 and -3 analysed 3 patients, when the number of descriptors is 36. That is each patient has a core set of descriptors reported, but not all patients, for example, have the same set of drugs administered.
Table 7
We have used the notation “on A → B” to indicate that an items which had been annotated initially by the clinician as an “A”, have since been reclassified by the expert as a “B”. If the item is followed by a “*” this implies that the changed annotation is now consistent with that predicted by the then-current version of the rule-set. (Remember that when the rule-set changes all the instances are re-evaluated against the revised rule-set.)

<table>
<thead>
<tr>
<th>Expert</th>
<th>Rule A</th>
<th>Rule B</th>
<th>Rule C</th>
<th>Rule D</th>
<th>Rule E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>157 A → B*</td>
<td>11 A → B</td>
<td>2 A → D*</td>
<td>2 A → D</td>
<td>3 data edits*</td>
</tr>
<tr>
<td>B</td>
<td>14 B → A*</td>
<td>2 A → C</td>
<td>7 A → C*</td>
<td>2 rule edits</td>
<td>3 A → U</td>
</tr>
<tr>
<td>C</td>
<td>12 C → B*</td>
<td>1 rule edit</td>
<td>1 B → U</td>
<td>1 B → D*</td>
<td>1 data edit</td>
</tr>
<tr>
<td>D</td>
<td>13 D → C</td>
<td>11 D → C*</td>
<td>1 C → E</td>
<td>2 rule edits</td>
<td>1 B → U</td>
</tr>
<tr>
<td>E</td>
<td>3 D → U</td>
<td>2 D → U</td>
<td>1 rule edit</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* To remove impermissible values.

Table 8
The rules associated with categories B, C, and D are also largely conjunctive, and tend to have values on a continuum from those associated with category “A” to those associated to category “E”. Fig. 3 shows an example of one of the confusion matrices created during the clinician interviews.

Clinician-1 followed roughly the refinement strategy suggested in Section 3; note there are no “E” rows in this confusion matrix which means that none of the instances classified by the expert as an “E” was classified as anything else by the rule-set. In fact the expert chose to consider cells (A, E) (B, E) (C, E) followed by (B, D) (D, B) and then (A, C) (A, B) (B, A) (D, E) (D, C) (B, C) (D, C) (B, D) (C, B) and (C, D). In the early stages of the analysis, very obvious inconsistencies were encountered and dealt with, and later it often became an issue of fine-tuning the rule-set or the dataset to achieve the classification which the expert wanted between two “adjacent” (1-distance apart) classifications. Table 7 gives a summary of the changes made to the “cells”. Here we provide an overview of the typical decisions made by this domain expert:

- **Inadmissible readings:** In cell (A, E), the expert considered that 3 of the values given in the dataset for heart rate were clearly inadmissible (values of 372, 7, 3); he changed those values to null values, and reclassified each of the cases as “unclassified” as he felt there was insufficient information to make a classification. He dealt with a further instance in cell (B, E) similarly.

- **Extrapolated data points:** Several times the expert agreed that the actual information provided in an instance was not sufficient to make a decision, and agreed, for several of the missing values, he had looked at the corresponding values in the immediately preceding and following time-periods and had effectively used extrapolated values when making his decisions. In all cases he agreed that the instances should have their classifications changed to “unclassified”. (This raises the issue of whether a further trending facility should be developed for INSIGHT and used with selected features.)

- **Significant values overlooked:** In many instances, e.g., cell (D, B), the expert agreed that the annotation should be changed as he had failed, when doing his initial classification, to note an important feature–value pair, in this case FiO2 values of 0.55.

- **Deciding borderline values:** In handling many of the “adjacent” cells where the distance between them is just one (e.g., cells such as (A, B) (B, C) (C, B)); the expert in some circumstances reclassified the instances, and in others he modified the appropriate rules to achieve his desired classification for the instances.

This expert made 297 changes to annotations. Note some annotations might well have been changed several times: for example, an instance originally annotated as an A, might initially be re-annotated as a “D”, and finally following a rule change, might be re-annotated as a “C”. In summary, this expert during the process of this refinement modified 51.6% (297/576) of his annotations: 272 changes (47.2%) were to adjust annotations which were on the borderline between two adjacent categories; 15 (2.6%) were “non-adjacent” annotations (these changes were often due to the expert overlooking a piece of information in the patient record which he accepted as important when it was brought to his attention (by INSIGHT)); and the remaining 10 changes (1.7%) were to reclassify instances as unclassified (often when he noted an unacceptable values for a “core” descriptors e.g., HR).

In this session, we observed two significant types of rule refinements, namely:

- **Adding a new rule, e.g., clinician-1 in phase 2 added a new conjunctive rule to category “E”: ADRENALINE (high) AND NORADRENALINE (high).**

- **Refining the conditions of a set of rules based on a common feature, say FiO2.** Note that all the values returned for FiO2 are effectively decimals (in the range 0.0–1.0); also note that all the ranges for FiO2 are continuous. Table 8 gives the values for FiO2 for a number of categories, both before and after refinement.
In the second (pilot) session, which lasted several hours, we started with the rule-set which had been produced in this first session (when the expert had processed the data associated with patient 705), and used that as the starting point to make the annotations of the remaining 9 patients (see Table 5) consistent with this rule-set or a variant of this rule-set. In this session the number of annotated instances to be dealt with was 2130 (i.e., 2706 – 576).

It should be noted that as a result of the changes made earlier to INSIGHT the progress in this session was considerably faster (on average operations took about a third of the time; some operations such as reclassifying blocks of instances are even quicker).

At the start of this session, the rule-set produced in phase 1 gave a 58.3% (1609 of 2760) agreement with the annotations created by the domain expert across all 10 patients; when the 135 unclassified instances are removed we get a 58.9% (1545 of 2625) agreement. By the end of the refinement session this agreement had increased to 96.4% (2663 of 2761), or when the 170 unclassified instances had been removed, to 100.0% (2591 of 2591) (with the corresponding kappa values being: 0.95 and 1.0). The expert initially chose to view the same parameters as he did at the end of the first session, but part way through he re-introduced dobutamine to the set. The strategy followed by the expert for refining these instances was very similar to that given above.

A summary of phase 2 is provided in Table 9. Given that the number of instances considered here is nearly four times as large as considered in phase 1, there are a relatively smaller number of changes, the exception being the number of instances which have been reclassified as “unclassified”. As noted before many instances are unclassified as “core” data elements are missing; clearly one is never going to capture all the data, but the expert noticed that data is often missing at critical points when patients experience a significant deterioration; this issue will be raised with nursing staff to see if the overall data collection rates can be improved. (We also noted earlier that data tends to be sparse when patients first come to the ICU and just before they are discharged.)

### 4.2. Study with clinician-2

In the session with clinician-2, which lasted about 2 h, we started with the rule-set which had been produced by clinician-1 as the result of reviewing all 10 patients. Clinician-2 had annotated three patient datasets, namely those of patients 708, 728, and 733, giving a total of 717 instances. (Clinician-1 annotated time points from 10 patients; the smaller number of 3 patients was chosen for subsequent clinicians to make the task more manageable.) This clinician also decided it was hard to review all the parameters reported for each time instance and chose, generally, to limit the ones he considered to adrenaline, blood pump speed, CVP, dobutamine, FI0₂, gelofusin, hartmanns, HR, LiDCO CI, MAP, noradrenaline, PiCCO derived parameters, propofol, sodium chloride, SpO₂, temperature, urine output, and vasopressin (18 parameters out of a set of 36 possible parameters).

The strategy followed by the expert for refining these instances was very similar to that used by clinician-1. At the start of this session the final rule-set produced by clinician-1 gave a 40.0% agreement with the annotations created by this domain expert for patient 708 (40 instances), and by the end of this session the agreement had increased to 97.5%. Table 10 gives results for the three patients; the percentage agreement, after dataset and rule refinements, for these instances is remarkably high: being 97.6% when unclassified cases are included and 99.6% when they are not. The corresponding kappa values before refinement (−0.02 and −0.02), with and without unclassified items,  

### Table 9

Summary of actions taken by clinician-1 in phase 2.

| Number of instances in the set | 2130 |
| Number of instances/annotations viewed | 225* |
| Number of data values edited/removed | 7 |
| Number of annotations changed to unclassified | 97 |
| Number of annotations left as “inconsistencies” | 16 |
| Number of annotations changed to another A–E level (excluding “unclassified”) | 1 |
| Number of changes to the rule-set | 6 |

* This figure is approximate as there are several ways in which it could be calculated.
Table 10
Summary of clinician-2’s refinement. (Each cell contains the percentage, the corresponding kappa value, and the number of instances matched against the total number of instances.)

<table>
<thead>
<tr>
<th></th>
<th>All 3 patients (before)</th>
<th>All 3 patients (after)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All instances considered</td>
<td>10.5%; −0.02</td>
<td>97.6%; 0.97</td>
</tr>
<tr>
<td></td>
<td>75/717</td>
<td>700/717</td>
</tr>
<tr>
<td>Unclassified instances excluded</td>
<td>10.6%; −0.02</td>
<td>99.6%; 0.99</td>
</tr>
<tr>
<td></td>
<td>75/710</td>
<td>700/703</td>
</tr>
</tbody>
</table>

Table 11
Summary of clinician-3’s refinement. (Each cell contains the percentage, the corresponding kappa value, and the number of instances matched against the total number of instances.)

<table>
<thead>
<tr>
<th></th>
<th>All 3 patients (before)</th>
<th>All 3 patients (after)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All instances considered</td>
<td>90.6%; 0.87</td>
<td>98.1%; 0.97</td>
</tr>
<tr>
<td></td>
<td>571/630</td>
<td>618/630</td>
</tr>
<tr>
<td>Unclassified instances excluded</td>
<td>90.0%; 0.88</td>
<td>98.4%; 0.98</td>
</tr>
<tr>
<td></td>
<td>568/625</td>
<td>615/625</td>
</tr>
</tbody>
</table>

are both classified as “no agreement”; after refinement the corresponding kappa values are 0.97 and 0.99 and so are both classified as “almost perfect agreement”.

4.3. Study with clinician-3

As clinician-1 and clinician-2 had worked closely together on the development of a patient scoring system it was thought important to have a further consultant from the ICU at Glasgow Royal Infirmary undertake the same analysis as clinician-2. So clinician-3 annotated the sets of instances for the same 3 patients as clinician-2 (717 instances); this annotation was done independently. Further, he took as his starting point the last rule-set produced by clinician-2. Once again he then used INSIGHT to help him make these 2 perspectives more consistent; and again he revised the dataset, the annotations of his dataset, and the rule-set. (In fact, this expert followed a very similar refinement strategy to the other 2 experts, and so it will not be described here.) Again as with the other clinicians, he found the full set of descriptors too large to handle, and in this case clinician-3 excluded 9 descriptors and so used the remaining 27. The initial agreement which clinician-3 achieved between his dataset (annotations) and the initial rule-set was 90.6% (571/630) if one includes unclassified instances in the calculation, and 90.9% (568/625) when one excludes them.11 As a result of the refinements clinician-3 made to both the dataset and to the rule-set, these figures became 98.1% (618/630) when one includes unclassified instances in the calculation, and 98.4% (615/625) when one excludes them. These results are summarized in Table 11; both the kappa values after refinement suggest that there is “almost perfect agreement”.

The changes by this expert essentially included refinements of the annotations, and rule refinements. Here we give a summary of the refinements made:

- The expert changed the annotation of 61 instances.
- 87 instances were accidentally deleted from the dataset (reducing the remaining dataset from 717 to 630).12 The expert had intended changing the category of these instances from C to D. (This information is included at a later stage of the analysis.)
- The expert changed the annotation of a further 16 instances.
- The expert then revised the rule-set.
- The expert changed the annotation of a further 60 instances.

The total number of annotations changed was 224 (including the deleted instances); below we summarize these amendments:

As noted earlier, the expert had intended reclassifying the deleted instances from category C to D; when these are included the changes to the annotations become:

- NULL to a category (1) [0.14%].
- Category to NULL (5) [0.70%].
- Changes involving adjacent categories (211) [29.4%].
- Changes involving non-adjacent categories (7) [1.0%].13

where the percentages given in [ ] are based on the original number of 717 instances. Once again we see that the majority of the changes are made to “adjacent” categories, and some feature–values have been deemed inadmissible and hence the instances have been reclassified as “unclassified”.

As noted above, clinician-3 made a number of changes to his annotations, and then refined 5 out of the 9 rules in the initial rule-set. (Note all the changes made are to the same descriptor, MAP.)

4.4. Comparisons between clinician-1, clinician-2, and clinician-3

Table 12 indicates that each of the 3 clinicians has achieved a high level of correlation between their individual annotations and own rule-set. Further clinician-1 and clinician-2 show a high degree of correlation (~96.5% including unclassified instances and ~96% if they are excluded); whereas the agreement with clinician-3 is consistently ~7% lower.14 Even these agreements are higher than we might have predicted. Just one of the kappa values in Table 12 is interpreted as “substantial agreement”; the remainder (17) are interpreted as “almost perfect agreement”. The next section discusses an approach to reconciling the perspectives of the 3 experts.

4.5. Attempting to achieve a consensus between the 3 experts: towards a methodology for creating consensuses between multiple experts

We can think of the instances reported by the patient management system as a vector for each time point. So suppose there are N features, then at time point t, we will have a vector of N feature–value (FV) pairs

\[
\langle FV_{1t}, FV_{2t}, \ldots, FV_{Nt} \rangle
\]

We shall refer to this as a “facts vector”. Moreover, if we are trying to make the classification consistent between 3 experts, say Ea, Eb, and Ec, then we also have their classifications for this instance intended changing the category of these instances from C to D. (This information is included at a later stage of the analysis.)

11 We explain further in the section why the total number of instances used in this analysis is 630 and not 717.

12 As this error occurred after about an hour into the refinement process, we felt we could not ask the clinician to restart the refinement process. Moreover, INSIGHT has now been modified so that the user is now asked to confirm deletions before they occur.

13 Again the expert noted that this was because he had overlooked important descriptor values when he had reviewed these instances initially.

14 If this approach indicates that Ea and Eb classify a particular instance (or set of instances) identically, we cannot infer that both experts have used identical descriptors to arrive at their classifications. They might, for instance, be focussing on different descriptors which are correlated in this particular instance. For example, Ea might focus on low BP, and Eb on abnormally high FiO2 values, and both the above descriptor–value pairs occur in this particular instance (time-slot). So this current approach identifies functional equivalence of decisions, but the later might well not be cognitively, or descriptively, equivalent (as indicated by the above example). If this becomes important, at some stage, the actual descriptors used by the several experts to make particular classifications, could be identified.
We distributed hardcopy versions of each of the clinicians’ final descriptor (\(\text{SpO}_2\)) mentioned in rule 1: we then systematically considered the remaining descriptors.

- There was a general discussion and agreed changes were written to a flip-chart. (See Section 4.6 for details.)
- These actual changes were then made to clinician-3’s final rule-set to produce a new “common” rule-set (common-4).
- INSIGHT was then used individually with each of the clinicians (clinician-1, clinician-2 and clinician-3) to refine their own dataset with respect to the common rule-set. In this stage of the study the experts were not allowed to change the (common) rule-set. The actual changes made to annotated datasets are reported, as are any significant comments and reservations.
- Further, INSIGHT calculated the agreements between each clinician’s refined dataset and the common rule-set.

We held a session with the 3 expert clinicians, with the analysts (DS and LM) as facilitators. As mentioned earlier we focused the discussion on the final rule-set produced by clinician-3. Specifically we discussed the role and values of each of the features occurring in the rules. Starting initially with \(\text{SpO}_2\) (the level of oxygen saturation in the patient’s blood), and systematically reviewing each of the features in each of the rules, Table 13 summaries the changes agreed by the clinicians. (As noted earlier this information was written to a flip chart which was visible to all the participants in the session.)

This represents a considerable number of changes. Some of the changes are simply fine-tuning the boundary between 2 categories (say B and C) to ensure that a particular instance is classified as the experts felt it should be. However the changes made to adrenaline’s values for most of the categories are quite considerable; in the previous rule-set the values assigned to both adrenaline and norepinephrine were comparable. Also 2 new “composite” rules have also been added; these both involve dobutamine.

The approach we chose to help resolve differences involved all the experts meeting and talking face-to-face in an attempt to achieve a greater level of agreement. On some occasions, changes suggested were readily accepted (for example the addition of “composite” rules involving dobutamine – as some composite rules were already part of the rule-set), on other occasions there were more detailed discussions. We appreciate that there are advantages in using the Delphi method [23] to resolve differences between experts but we chose not to use that approach as we felt that neither of the facilitators had enough medical knowledge to fulfill the mediator role within the Delphi approach adequately. Hripcsak and colleagues [42] have similarly attempted to resolve differences between experts; they have attempted to understand the several experts’ position on an issue, and then to mediate between them. But as noted above the facilitators did not, in this study, play that type of role.

### 4.6. Creating a common rule-set

After extensive discussions we decided, in the first instance, to adopt a simplified approach in which we first sought to create a common rule-set (see Fig. 4). This process is summarized below:

- We distributed hardcopy versions of each of the clinicians’ final rule-sets; and in fact we concentrated our discussions on that produced by clinician-3. We considered, across all categories, the first

- The experts would review the composite vector produced by INSIGHT.
- They would then decide on changes to be made to the values in the “fact vector”, and changes to both their annotations and rule-sets.
- The experts would then make the agreed changes to their own dataset and rule-set.
- INSIGHT would produce a further composite vector which highlights the remaining differences.

This cycle continues until the experts reach complete agreement, or they agree they are unable to make any further progress. In this latter case it is important that the experts and INSIGHT record their differences. A metric which evaluates the goodness of fit of each set of vectors would greatly help the above process, as without it, it might be hard to decide whether a proposed set of changes would be convergent or divergent. Additionally, it should be noticed that the whole process would become more complicated as the number of experts increases.

### Table 12

Comparison between final datasets and rule-sets for clinician-1, clinician-2 and clinician-3 on instances from three patients. Asterisked results correspond to analyses where the unclassified instances are removed from the calculation.

<table>
<thead>
<tr>
<th>Clinician-1’s final dataset</th>
<th>Clinician-2’s final dataset</th>
<th>Clinician-3’s final dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinician-1’s final rule-set</td>
<td>96.7%; 0.95 (693/717)</td>
<td>94.0%; 0.92 (674/717)</td>
</tr>
<tr>
<td>Clinician-2’s final rule-set</td>
<td>93.4%; 0.91 (670/717)</td>
<td>97.6%; 0.97 (700/717)</td>
</tr>
<tr>
<td>Clinician-3’s final rule-set</td>
<td>84.4%; 0.79 (605/717)</td>
<td>88.8%; 0.85 (637/717)</td>
</tr>
<tr>
<td></td>
<td>87.13%; 0.82 (587/674)</td>
<td>90.6%; 0.87 (637/703)</td>
</tr>
</tbody>
</table>

15 Also the focus of this study was essentially on how effective the INSIGHT tool and methodology are in resolving inconsistencies between 2 perspectives held by a single expert.
Fig. 4. Flow chart of a process to help experts make their several classifications more consistent.

Each clinician then attempted (using INSIGHT) to make his dataset consistent with the newly agreed (common) rule-set. The changes made by the several clinicians are listed below:

4.6.1. Clinician-1

B → D (1 change).\(^{16}\) Note this involves categories which are 2 “distances” apart. All the remaining refinements involve categories which are only 1 distance apart: B → A (53), B → C (14), C → D (12), D → E (22), and E → D (3). That is 105 changes in all.

Disagreement with classification suggested by the common rule-set: This clinician had annotated a particular instance as B and the common rule-set suggested a D (because of the MAP); the clinician said he thought this was a ‘bit harsh’ “because everything else in that instance was OK”. This occurred in total 6 times. This clinician disagreed with the common rule-set’s classification, and as a result he chose not to modify his annotations.

4.6.2. Clinician-2

A → NULL (3).

A → C (2) and B → D (6) changes. (Note this involves categories which are 2 “distances” apart.) All the remaining refinements involve categories which are only 1 distance apart: A → B (5), B → A (53), B → C (23), C → D (17) and D → E (49).

That is 158 changes in all.

No disagreements with the common rule-set were recorded.

4.6.3. Clinician-3

All the refinements involve which categories are only 1 distance apart:

A → B (5), B → A (35), B → C (4), C → B (5), C → D (6), D → C (8), and

D → E (49).

That is 112 changes in all.

No disagreements with the common rule-set were recorded.

Table 14 gives the percentage agreements between the common rule-set and the initial and revised datasets for each of the clinicians. The results for the final datasets imply that if we exclude unclassified instances from the calculations then there is perfect agreement between clinician-2 and clinician-3, and a very high agreement with clinician-1 (96.4%). With unclassified instances included these agreements reduce by 3.2%, 1.5%, and 0.3% respectively for the three experts. The (6) corresponding kappa values are interpreted as “almost perfect agreement”.

Moreover, the percentage changes between the initial and revised datasets for the 3 clinicians (when the unclassified

---

\(^{16}\) The notation used here is: ORIGINAL-Category→REVISED-Category (number-of-items-changed).
Noradrenaline and dobutamine at the D-level.

- Adrenaline and dobutamine at the D-level.
- Adrenaline and noradrenaline at the D-level.
- FiO2 (at D level) and dobutamine (at D level).
- FiO2 (at D level) and noradrenaline (at D level).
- FiO2 (at D level) and adrenaline (at D level).

The unclassified instances are removed from the calculation.

Comparison between the final common rule-set and the initial and final datasets for clinician-1, clinician-2, and clinician-3. Asterisked results correspond to analyses when unclassified instances are excluded from the calculations.

Agreement achieved are “almost perfect agreement” (Table 14). The high level of agreement between the clinicians, we believe, is explained by the practice, certainly at Glasgow Royal Infirmary and at many other UK ICUs, in which care is highly protocolized. Further care of individual patients is often shared between several consultants, and ward rounds are carried out jointly.

What is of considerable interest is the number and nature of the changes made (to the final rule-set of clinician-3) when all 3 clinicians met to determine a common rule-set. These changes effectively fell in to 3 categories:

(a) Changes to fine tune adjacent categories (e.g., categories B–C, and D–E).
(b) Substantial changes in the ranges recorded for a drug (in this instance adrenaline) across all 5 categories (A–E).
(c) Additional composite rules for category E.

Using this common rule-set, once the 3 clinicians had revised their datasets to be consistent with it, we again obtained some good agreements (96.4%, 100%, 100%) if one excludes the unclassified instances, with the corresponding kappa values indicating that the agreements achieved are “almost perfect agreement” (Table 14).

The crucial question now is how stable is the common rule-set? That is if we were to involve several further clinicians in the expert-consensus process how extensive would the changes be? Clearly this is very much an empirical question; see the Section 6 for discussion of our plans. However, let us hazard a guess as to what we might encounter. Earlier in this section, we introduced a classification for the sorts of changes which such a process has introduced. We anticipate that fine-tunings of type (a) will continue to be seen (no one is too concerned about this). Revising the ranges for drugs (as adrenaline above) to cover the several categories, we would expect to be pretty unusual. We would however, expect to encounter the inclusion of more “composite rules” as the clinicians become progressively more familiar with this approach (i.e., making explicit the rules which underlie their own decision-making processes). Indeed, composite rules so far have only appeared for

Table 12 reports agreements between clinician-1, clinician-2, and clinician-3 based on the annotation of 3 patients, where each clinician had annotated the instance set independently, and had developed his own rule-set. (To be more precise, clinician-2’s initial rule-set was based on the final one developed by clinician-1, and clinician-3’s initial rule-set had been based on that from clinician-2.) Each expert achieved a high degree of (intra-rater) correlation between his final set of annotations and his final rule-set; an average of 99.3% if one excludes unclassified instances. The agreement between clinician-1 and clinician-2 is better than those between clinician-3 and the other 2; the latter agreements being 87.1% and 90.6% between clinician-3’s final rule-set and the final datasets of clinician-1 and clinician-2 respectively (when unclassified instances are excluded). As reported in Section 4.4, just 1 of the kappa values in Table 12 is interpreted as “substantial agreement”; the remainder (17) are interpreted as “almost perfect agreement”.

Table 13 lists of the changes made to clinician-3’s final rule-set to create the common rule-set (common-4).

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>Category</th>
<th>Before refinement</th>
<th>After refinement</th>
</tr>
</thead>
<tbody>
<tr>
<td>SpO2</td>
<td>E</td>
<td>0–88</td>
<td>0–85</td>
</tr>
<tr>
<td>SpO2</td>
<td>D</td>
<td>89–91</td>
<td>86–91</td>
</tr>
<tr>
<td>FiO2</td>
<td>E</td>
<td>0.94–1.0</td>
<td>0.9–1.0</td>
</tr>
<tr>
<td>FiO2</td>
<td>D</td>
<td>0.70–0.93</td>
<td>0.70–0.89</td>
</tr>
<tr>
<td>HR (lower)</td>
<td>E</td>
<td>0.0–42.0</td>
<td>0.0–40.0</td>
</tr>
<tr>
<td>HR (lower)</td>
<td>D</td>
<td>43.0–45.0</td>
<td>41.0–45.0</td>
</tr>
<tr>
<td>HR (lower)</td>
<td>C</td>
<td>46.0–49.0</td>
<td>46.0–50.0</td>
</tr>
<tr>
<td>HR (lower)</td>
<td>B</td>
<td>50.0–59.0</td>
<td>51.0–55.0</td>
</tr>
<tr>
<td>HR</td>
<td>A</td>
<td>60.0–83.0</td>
<td>56.0–89.0</td>
</tr>
<tr>
<td>HR (upper)</td>
<td>E</td>
<td>141.0–500.0</td>
<td>141.0–300.0</td>
</tr>
<tr>
<td>HR (upper)</td>
<td>B</td>
<td>84.0–99.0</td>
<td>90.0–99.0</td>
</tr>
<tr>
<td>HR (upper)</td>
<td>A</td>
<td>60.0–83.0</td>
<td>60.0–89.0</td>
</tr>
<tr>
<td>Mean/MAP (upper)</td>
<td>D</td>
<td>110–129</td>
<td>120.0–129.0</td>
</tr>
<tr>
<td>Mean/MAP (upper)</td>
<td>C</td>
<td>100–109</td>
<td>110.0–119.0</td>
</tr>
<tr>
<td>Mean/MAP (upper)</td>
<td>B</td>
<td>91–99</td>
<td>100–109</td>
</tr>
<tr>
<td>Mean/MAP (upper)</td>
<td>A</td>
<td>71–90</td>
<td>71–99</td>
</tr>
<tr>
<td>Noradrenaline</td>
<td>E</td>
<td>2.5–10.0</td>
<td>2.0–10.0</td>
</tr>
<tr>
<td>Noradrenaline</td>
<td>D</td>
<td>1.8–2.4</td>
<td>1.0–1.9 (9)</td>
</tr>
<tr>
<td>Noradrenaline</td>
<td>C</td>
<td>1.0–1.7</td>
<td>0.50–0.9 (9)</td>
</tr>
<tr>
<td>Noradrenaline</td>
<td>B</td>
<td>0.1–0.9</td>
<td>0.10–0.4 (9)</td>
</tr>
<tr>
<td>Adrenaline</td>
<td>E</td>
<td>2.50–10.0</td>
<td>1.0–10.0</td>
</tr>
<tr>
<td>Adrenaline</td>
<td>D</td>
<td>1.80–2.40</td>
<td>0.5–0.9(9)</td>
</tr>
<tr>
<td>Adrenaline</td>
<td>C</td>
<td>1.00–1.70</td>
<td>0.3–0.4(9)</td>
</tr>
<tr>
<td>Adrenaline</td>
<td>B</td>
<td>0.10–0.90</td>
<td>0.05–0.2(9)</td>
</tr>
<tr>
<td>Dobutamine</td>
<td>E</td>
<td>42.1–200</td>
<td>61–200</td>
</tr>
<tr>
<td>Dobutamine</td>
<td>D</td>
<td>25.1–42.0</td>
<td>41–60</td>
</tr>
<tr>
<td>Dobutamine</td>
<td>C</td>
<td>10.6–25.0</td>
<td>21–40</td>
</tr>
<tr>
<td>Dobutamine</td>
<td>B</td>
<td>0.1–10.5</td>
<td>0.1–20</td>
</tr>
</tbody>
</table>

Additional category “E” rules: Note the first 4 rules below are structurally the same as the earlier set, but note that they are effectively different as the D-levels have been refined for all 3 of these drugs. And of course the last 2 rules are new (rules involving dobutamine and a further inotrope).

- FiO2 (at D level) and adrenaline (at D level).
- FiO2 (at D level) and noradrenaline (at D level).
- FiO2 (at D level) and dobutamine (at D level).
- Adrenaline and noradrenaline at the D-level.
- Adrenaline and dobutamine at the D-level.
- Noradrenaline and dobutamine at the D-level.

instances are excluded from the calculations) are 15.5%, 55.5% and 17.9% respectively. The changes to the corresponding kappa values are equally significant; for example, for clinician-2 the significance of the kappa values changed from “fair agreement” to “perfect agreement”. Thus the experts (with support from INSIGHT) have achieved a high level of agreement on these classification tasks.

Table 12 reports agreements between clinician-1, clinician-2, and clinician-3 based on the annotation of 3 patients, where each clinician had annotated the instance set independently, and had developed his own rule-set. (To be more precise, clinician-2’s initial rule-set was based on the final one developed by clinician-1, and clinician-3’s initial rule-set had been based on that from clinician-2.) Each expert achieved a high degree of (intra-rater) correlation between his final set of annotations and his final rule-set; an average of 99.3% if one excludes unclassified instances. The agreement between clinician-1 and clinician-2 is better than those between clinician-3 and the other 2; the latter agreements being 87.1% and 90.6% between clinician-3’s final rule-set and the final datasets of clinician-1 and clinician-2 respectively (when unclassified instances are excluded). As reported in Section 4.4, just 1 of the kappa values in Table 12 is interpreted as “substantial agreement”; the remainder (17) are interpreted as “almost perfect agreement”.

The high level of agreement between the clinicians, we believe, is explained by the practice, certainly at Glasgow Royal Infirmary and at many other UK ICUs, in which care is highly protocolized. Further care of individual patients is often shared between several consultants, and ward rounds are carried out jointly.

What is of considerable interest is the number and nature of the changes made (to the final rule-set of clinician-3) when all 3 clinicians met to determine a common rule-set. These changes effectively fell in to 3 categories:

(a) Changes to fine tune adjacent categories (e.g., categories B–C, and D–E).
(b) Substantial changes in the ranges recorded for a drug (in this instance adrenaline) across all 5 categories (A–E).
(c) Additional composite rules for category E.

Using this common rule-set, once the 3 clinicians had revised their datasets to be consistent with it, we again obtained some good agreements (96.4%, 100%, 100%) if one excludes the unclassified instances, with the corresponding kappa values indicating that the agreements achieved are “almost perfect agreement” (Table 14).

The crucial question now is how stable is the common rule-set? That is if we were to involve several further clinicians in the expert-consensus process how extensive would the changes be? Clearly this is very much an empirical question; see the Section 6 for discussion of our plans. However, let us hazard a guess as to what we might encounter. Earlier in this section, we introduced a classification for the sorts of changes which such a process has introduced. We anticipate that fine-tunings of type (a) will continue to be seen (no one is too concerned about this). Revising the ranges for drugs (as adrenaline above) to cover the several categories, we would expect to be pretty unusual. We would however, expect to encounter the inclusion of more “composite rules” as the clinicians become progressively more familiar with this approach (i.e., making explicit the rules which underlie their own decision-making processes). Indeed, composite rules so far have only appeared for
category E, but we should expect them also to occur in categories D, C and B; and possibly category A.

5. Contributions of this work

We attempt in this section to summarize the main contributions of this study. Firstly, we have demonstrated the benefits of experts being able to see their knowledge applied on a series of relevant tasks, and being able, when discrepancies exist, to refine knowledge in the context of a specific task. Secondly, we have produced a useful tool to help an expert appreciate how two perspectives on the same task are inconsistent; further this tool allows the expert to explore ways in which the two knowledge sources can be made (more) consistent. Thirdly, we confirmed the advantage, in some circumstance, of a simple information checking system as opposed to a more complex system which is able to (semi)-automatically extract the knowledge from a set of labelled instances. Fourthly, an important “side” effect of using INSIGHT with domain experts is that the datasets are tidied up and extreme values and wrong classifications are often identified and modified (a very useful pre-processing step for machine learning algorithms). Fifthly, we confirmed the need, when acquiring knowledge from domain experts, to determine whether a particular category has sub-categories and if so to get the expert to articulate them. Finally, we confirmed the need to have a domain expert critically review, whenever possible, any rules (knowledge) produced by an automated system.

6. Further work

Future work has been grouped under 2 headings, namely: ICU scoring systems and modifications and further applications of INSIGHT.

6.1. ICU scoring systems

Firstly, at least one clinician has indicated that the rate of change of parameters is a factor which he takes into account under some circumstances; we plan to investigate adding this functionality to INSIGHT. Secondly, we plan to evaluate the ICU scoring system across several ICUs and with a larger number of experts. The central task which INSIGHT has been used to investigate, to date, is the development of a reliable patient scoring system. So far, we have applied INSIGHT to data from only 3 patients from a single ICU, and this information has been evaluated by just three domain experts (clinician-1 annotated 10 patients). Clearly, if the scoring system is to be used widely it will need to be evaluated with a larger and more disparate group of patients and with considerably more domain experts. This evaluation is currently being planned; this evaluation will hopefully address a number of the shortcomings noted in Section 4, including using a range of different approaches to obtain (greater) consensus between experts with differing views.

6.2. Modifications and further applications of INSIGHT

Firstly, we plan to use INSIGHT with a range of other tasks including the classification of botanical species and other clinical diseases. In many situations, experts find it hard to articulate the actual distinctions between different categories; as the next point indicates, INSIGHT should help with this process. Secondly, we plan to investigate INSIGHT’s mode to create rules from instances. We noted in Section 3.2 that INSIGHT had such a mode, and that to date it had not been used by domain experts on a range of demanding real-world tasks. Clearly, we believe that this mode will be valuable for domain experts who will then not need to create a set of rules. We plan to use this facility with a number of domains and with a range of experts.

Thirdly, a thorough investigation of the differences between expert-provided rule-sets versus those inferred by several machine learning algorithms from the same datasets (noisy and INSIGHT-refined) is planned. Fourthly, we are planning to develop a variant of INSIGHT which can be applied to planning (synthetic) tasks. This will be more demanding than for classification tasks, but we believe it is possible, and moreover that it would be a useful additional tool in assessing expertise.

Conflict of interest

No conflicts to report.

Acknowledgements

We are grateful to Sunil Sharma and Mark Winter who implemented the earlier versions of REFINER. Andy Aiken and Laura Moss were both partially supported by EPSRC Studentships and the AKT project when they undertook aspects of this work. David Corsar (Aberdeen) provided helpful comments on various aspects of the analysis. Finally, we wish to acknowledge Kathryn Henderson and Jennifer McCallum (CareVue Project, NHS Glasgow) for their support in providing patient datasets and the staff of the ICU Glasgow Royal Infirmary for helpful discussions about this project.

Appendix A. High level summary of qualitative assessments

Below we give outline descriptions for each of the 5 categories, where E corresponds to the most severely ill patients:

A. Patient’s cardiovascular system (CVS) normal, with no adrenaline or noradrenaline and low levels of oxygen; urine production often essentially normal (or is well established on renal replacement therapy).

B. Patient CVS nearly normal, probably needs low levels of adrenaline or noradrenaline and oxygen.

C. Patient CVS system is effectively stable; probably on moderate dosages of adrenaline or noradrenaline and oxygen.

D. Patient’s CVS system is moderately unstable and/or on high doses of adrenaline or noradrenaline or fluid to retain stability.

Most parameters suggest the time-slot is in category A or B, but if any of the following conditions are met, then it should be assigned to category C:
- heart rate: moderately low or moderately high;
- MAP: moderately low or moderately high;
- adrenaline: moderate dose;
- noradrenaline: moderate dose;
- FiO2: moderate;
- SpO2: moderately low.

D. Patient’s CVS system is moderately unstable and/or on high doses of adrenaline or noradrenaline or fluid to retain stability.

Most parameters suggest the time-slot is in category A or B, but if any of the following conditions are met, then it should be assigned to category D:
- heart rate: low or high;
- MAP: low or high;
- adrenaline: high dose;
- noradrenaline: high dose;
- FiO2: high;
- SpO2: low.

E. Patient’s CVS is very unstable (which is usually true in early phases of resuscitation, or following a new event) with low BP and high HR or rapidly changing adrenaline or noradrenaline dosage, and requires substantial fluid inputs.
Most parameters suggest the time-slot is in category A or B, but if any of the following conditions are met, then it should be assigned to category E:
- heart rate: extremely low or extremely high;
- MAP: extremely low or extremely high;
- adrenaline: extremely high dose;
- noradrenaline: extremely high dose;
- FiO2: extremely high;
- SpO2: extremely low;

References