

A system to detect inconsistencies between a domain expert's different perspectives on (classification) tasks

Derek Sleeman¹, Andy Aiken¹, Laura Moss¹, John Kinsella², & Malcolm Sim²

¹ Department of Computing Science, University of Aberdeen, Aberdeen, AB24 3UE; {[d.sleeman](mailto:d.sleeman@abdn.ac.uk), [a.aiken](mailto:a.aiken@abdn.ac.uk), [lmoss](mailto:lmoss@abdn.ac.uk)}@abdn.ac.uk

² Department of Anaesthesia, Glasgow Royal Infirmary, University of Glasgow, Glasgow, G31 2ER; {[j.kinsella](mailto:j.kinsella@clinmed.gla.ac.uk)}@clinmed.gla.ac.uk

ABSTRACT

This paper discusses the range of knowledge acquisition, including machine learning, approaches used to develop knowledge bases for Intelligent Systems. Specifically, this paper focuses on developing techniques which enable an expert to detect inconsistencies in 2 (or more) perspectives that the expert might have on the same (classification) task. Further, the INSIGHT system has been developed to provide a tool which supports domain experts exploring, and removing, the inconsistencies in their conceptualization of a task. We report here a study of Intensive Care physicians reconciling 2 perspectives on their patients. The high level task which the physicians had set themselves was to classify, on a 5 point scale (A-E), the hourly reports produced by the Unit's patient management system. 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.

Inconsistencies between these 2 perspectives are displayed as a confusion matrix; moreover INSIGHT then allows the expert to revise both the annotated datasets (correcting data errors, and/or changing the assigned categories) and the actual rule-set. Each expert achieved a very high degree of consensus between his refined knowledge sources (i.e., annotated hourly patient records and the rule-set). Further, the consensus between the 2 experts was ~95%. The paper concludes by outlining some of the follow-up studies planned with both INSIGHT and this general approach.

1: Introduction

Contemporary knowledge-based systems, as their expert systems predecessors (Buchanan & Shortliffe 1984), 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, (Corsar & Sleeman 2007). This paper surveys 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 (PS), the use of machine learning in knowledge acquisition / capture, as well as more recent attempts to infer information from data sets produced by large numbers of users of systems like Open Mind, (Singh et al, 2002).

A central problem, since the inception of Expert Systems, is how to deal with the uncertainty inherent in such knowledge bases (Buchanan & Shortliffe, 1984). EMYCIN (Buchanan & Shortliffe, 1984) associated certainty factors with particular pieces of information (both facts & rules) and evolved a calculus which allows the uncertainty associated with decisions to be calculated, and then reported to the user. 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, (Pearl, 1988). Both these approaches provide pragmatic approaches to the handling of uncertainty associated with expertise. Clearly, however there are different types of uncertainty associated with pieces 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 Psychology & Artificial Intelligence.

The focus of the work reported here is an attempt to get experts to provide 2 perspectives on a classification task, and then to provide a system / tool 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). 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 up to 60 pieces of information) on a 5-point A-E scale where E is severely ill.

The rest of the paper is structured as follows: section 2 gives an overview of ICU patient management systems, and the types of information which they produce; additionally, patient scoring systems are discussed. Section 3 gives an overview of the cognitive science literature on expertise, on knowledge acquisition / capture including the important role which machine learning has played in these activities; thirdly we review cooperative knowledge acquisition and knowledge refinement systems. Section 4 provides a conceptual overview of the INSIGHT system which takes 2 perspectives on an expert's classification knowledge, detects inconsisten-

cies 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 5 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 by both experts. Section 6 outlines several of the contributions of this work. Section 7 concludes the paper by outlining some planned follow-up studies.

2. Overview of Patient Management Systems used in Intensive Care Units (ICUs)

This section gives an overview of patient monitoring systems which are used in Intensive Care Units (ICUs), together with examples of parameters collected. We also discuss the need for patient scoring systems, and outline a 5-point qualitative scale which we have developed.

Many ICUs have patient management systems which collect the patients' physiological parameters, records nursing activities, and other interventions (such as the administration of drugs and boluses of fluids). This information is typically collected at specified time periods say every minute or hour, is recorded on a data base 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 data sets. This is the focus of an aspect of our work which has led us to produce the ACHE (Architecture for Clinical Hypothesis Evaluation) infrastructure, (Moss et al., 2008). 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 (GRI), a Philips CareVue, records up to 60 parameters. Table 1 lists the principal parameters, and lists the frequency of recording in the current data set. It should be noted that the data sets 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 & different analyses are performed on the data "off-line".

2.1 Patient Scoring System

For a variety of reasons it would be helpful to clinicians if they were able to obtain a regular summary of each patient's overall condition. Such information would be useful to determine whether there has been any appreciable progress / 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-2 scale

(Knaus *et al.*, 1991) is widely used in ICUs in the western world, but the APACHE score is created only once during a patient's ICU stay, usually 24 hours 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.¹

Parameter	Recorded Interval
Heart Rate	Hourly
Temperature	Hourly
Mean Arterial Pressure (MAP)	Hourly
Diastolic	Hourly
Systolic	Hourly
FiO ₂	Hourly
SpO ₂	Hourly
Urine output	Hourly
Central Venous Pressure (CVP)	Hourly
LiDCO (If applicable to patient)	Hourly
Drug Infusions (eg Adrenaline, Noradrenaline)	As applicable
Fluid Infusions	As applicable
Dialysis Sessions	As applicable

Table 1: Parameters used in the study

The clinical authors of this paper (JK & MS) have been addressing this issue for some while. More recently we have produced a 5-point (high-level) qualitative description of ICU patients, which can be summarized as follows:

- E** Patient is highly unstable with say a number of his physiological parameters (e.g., blood pressure, heart rate) having extreme values (either low or high).
- D** Patient more stable than patients in category E but is likely to be receiving considerable amounts of support (e.g., fluid boluses, drugs such as Adrenaline, & possible high doses 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 & 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 doses of inspired oxygen

For more details on the descriptions, please see Appendix A.

¹ Adrenaline normally raises a patient's blood pressure through its inotropic effect.

The objective of the study 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 top-level outline of the study is:

- 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 1-23 days)
- One of the clinical investigators (MS) 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 Philips CareVue system i.e., that contained in the spreadsheets
- Further we asked the same clinician to articulate rules to describe each of the 5 categories, (i.e., A-E).
- We used the INSIGHT tool (described below) to help this clinician make this data set & his rule set more consistent by modifying, as he saw fit, either the annotations in the data set, his rule-set, or both.
- The second clinician (JK) annotated 3 of the patients' data sets, again using the same qualitative scale (A-E)
- We used the INSIGHT tool to help the second clinician (JK) make his data set consistent with the rule set produced by the first clinician. This clinician was, of course, allowed to modify both his annotations of the data set & the actual rules).

More details of this study are given in section 5.

3. Literature Overview

This section gives a Cognitive Science perspective on the acquisition of expertise (section 3.1), provides an overview of knowledge acquisition (including machine learning) approaches in section 3.2, and discusses cooperative knowledge acquisition and knowledge refinement systems in section 3.3.

3.1 The Cognitive Science Perspective on the acquisition of Expertise

The classic book on Protocol Analysis by Ericsson & Simon (1993) 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*. In the case of protocol analysis they further argue that the process of verbalizing the steps of problem solving does not perturb the expert's actual problem solving processes. Effectively, Ericsson & Simon introduced the distinction between "active" knowledge which is used to solve tasks as opposed to "passive" knowledge which is used to discuss tasks / a domain.

This has been a recurrent theme / perspective in much of cognitive science and in the study of expertise since that time, as is illustrated by the very nice study re-

ported by Johnson (1983). This investigator attended a medical professor's lectures on diagnosis where he explained the process. The investigator then accompanies the professor's ward round (with a group of medical students) & noticed a difference in his procedures. When challenged about these differences the medical 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 expertise / knowledge acquisition (KA) is that one should ask an expert to solve specific task(s), and (preferably) explain what s/he 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).

3.2 Summary of Knowledge Acquisition including uses of Machine Learning to extract domain knowledge in a number of domains

In a recent overview at the K-CAP 2007 conference, Sleeman (Sleeman et al, in press) argued that Knowledge Acquisition (KA) is “a broad church” and consists of a very wide range of approaches including:

- Interviewing of domain experts by Knowledge Engineers: an approach which was dominant in the early development of Expert Systems (Buchanan & Shortliffe, 1984).
- 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, (Diaper, 1989).
- Problem-Solving Method (PSM) driven systems such as MOLE, MORE, SALT acquire more focussed information which is sufficient to satisfy a *particular* type of problem solver / PSM. The use of these systems is less demanding for the domain expert as the information collected is generally less, and the purpose of the information collected is usually more apparent, (Marcus & McDermott, 1989.)
- Machine-learning approaches have played an important role of transforming sets of usually labelled instances into knowledge (usually rule sets). Given the context of this volume I provide some more detail of these approaches below.
- 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 & have extracted useful information (Etzioni et al, 2005).
- Capitalizing on greater connectivity & 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 to extract some very interesting concepts & associations (Singh et al, 2002). Similarly, von Ahn has exploited peoples' enthusiasm for on-line game playing, von Ahn (2006).

Michalski and Chilausky (1980) had a notable early success using Machine Learning approaches to extracting knowledge / rules from instances, in the domain of crop disease. The soya bean crop is of major importance to the state of Illinois, and so it employed a number of plant pathologists to advise farmers on crop diseases. Michalski and Chilausky studied the standard reference book on the subject, and also spent 40 or so hours interviewing an expert. This allowed them to determine, what they believed were, the appropriate set of descriptors for soya bean diseases. Subsequently, they developed a questionnaire to illicit, from farmers, examples of actual crop diseases which they had experienced; in fact, they obtained nearly 700 such cases. They then trained a version of the ID3 program (Quinlan, 1986) with 307 instances, and used the trained system to classify 376 test cases. The performance of the trained program was impressive; it only misclassified 2 instances whereas humans following the information given in the standard text-book misclassified 17% of the cases. A final step in this project was to extract a series of IF-THEN-ELSE rules from the ID3 tree, for day-to-day use by the plant pathologists and farmers.

For a more recent survey of the application of Machine Learning approaches to real-world tasks, see Langley & Simon (1995).

3.3 Cooperative Knowledge Acquisition & Knowledge Refinement systems

Building large knowledge bases is a demanding task; and particularly so if one is working in a domain where the knowledge / information is still "fluid". When one attempts to use such knowledge bases in conjunction with an appropriate inference engine to solve real-world tasks, one often finds that information is missing (and hence needs to be acquired), or the system gives answers to tasks which the domain expert says are incorrect (and hence the knowledge base needs to be refined). Again, if the domain is at the cutting edge of human knowledge then it is not possible to draw on an existing source of knowledge to support the processes of acquisition and refinement noted above, and hence one must use a well-chosen domain expert to act as the oracle. For obvious reasons the systems which have been built, by our group and others, to fulfil this role are often referred to as *Cooperative Knowledge Acquisition and Knowledge Refinement systems*. See Sleeman (1994) for a review of such systems. Over the last decade or more we have implemented systems which are able to refine knowledge bases (KBs) in a variety of formalisms including rules, cases, taxonomies, and causal graphs. The family of systems which are most relevant to this discussion are those which are able to refine cases, and they are discussed in the next sub-section.

3.3.1 The REFINER Systems

The REFINER family of programs have been designed to detect inconsistencies in a set of labelled cases. That is, these systems are provided with a set of categories which the domain expert believes are relevant to the domain, a set of descriptors needed to describe the domain, and a set of labelled cases / instances. The descriptors can be of a variety of types including real, integer, string and hierarchical. If the latter, then the system requires some further information about the nature of the taxonomy (for example, Figure 1). Table 2 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 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 3 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.



Figure 1: The hierarchy for the Disease descriptor

Case	Heart Rate (HR)	Diastolic Blood Pressure (DBP)	Disease	Category
1	50	90	Disease 1	A
2	56	90	Disease 2	A
3	52	101	Disease 3	A
4	50	95	Disease 3	B
5	56	97	Disease 3	B
6		89	Disease 5	A
7	52	97	Disease 3	B

Table 2: Sample dataset used to illustrate Refiner++

Category	HR	DBP	Disease
A	50 – 56	89 – 101	Any Disease
B	50 – 56	95 – 97	Disease 3

Table 3: The category descriptions generated by Refiner++

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 (due perhaps to a typing error)
- Reclassifying a case / instance
- Shelving a case to work on it subsequently
- Adding an additional descriptor to all the cases
- Creating a disjunction by excluding a value or range of values from a category description

Considering the dataset shown in Table 2, the category descriptions are inconsistent (a case with a DBP value in the range 95 – 97 and a Disease value of Disease 3 could not be unambiguously categorised) 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 DBP in case 4 to 97
- 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 4 shows the updated (non-overlapping) category.

Category	HR	DBP	Disease
A	50 – 56	89 – 101, except 95 – 97	Any Disease
B	50 – 56	95 – 97	Disease 3

Table 4: Updated category descriptions

We have so far effectively implemented 3 systems:

- **REFINER** (Sharma & Sleeman, 1988) was the first system; it was incremental in that it processed a single case / instance and attempted at each stage to remove any inconsistencies detected.
- **REFINER+**: 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 data set could be overwhelming for the expert. To help contain the situation we evolved several heuristics namely:

- A change which removes a considerable number of inconsistencies is preferred over one which removes a smaller number of inconsistencies;
- A change which makes a smaller number of changes to the data set is preferred over one which makes more extensive changes
- **REFINER DA:** The essential difference between REFINER DA & its predecessor REFINER+ is that it combined aspects of the two earlier systems. Namely the domain expert is asked to suggest several cases which he/she thought were prototypical of the several categories, from which descriptions of the several categories were inferred as described above. Then this version of REFINER attempted to cover additional cases without causing the set of category descriptors to become inconsistent.

3.3.2 Critique of REFINER DA & effectively the REFINER family of systems

The machine learning algorithm attempts to create, in each version of REFINER, a set of non-overlapping descriptions for the categories; moreover, each of the descriptors 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 in guiding this selection of features. So in principle the feature-value pair SpO₂ (96-100) could be used to determine that a patient was in category A (i.e., dischargeable), whereas if that same case had a further feature-value pair of FiO₂: (95-100), this would be clinically described as a very sick patient. So from working with REFINER DA with this data set we made two important observations:

- The feature-value pairs chosen to make a category distinct 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 the decision trees created by C4.5)
- An expert might effectively sub-divide a category like E into a number of sub-categories, which he might not initially articulate. (That is a patient can be in category E for one of several *distinct* reasons: e.g., poor heart rate, or poor oxygen saturation.) If the domain expert does not articulate these sub-categories then category E will be an amorphous category which will influence the descriptions inferred and this in turn will affect the other categories inferred by REFINER. Additionally if the sub-categories are articulated then it is likely that there will only be a small number of examples in each of the sub-categories, which again will mean that the machine learning algorithms will have difficulties in extracting domain-relevant descriptions.

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

4. Conceptual Design of INSIGHT

Below we give the design criteria for a system, INSIGHT, which we believe addresses (some of) the difficulties noted at the end of the last section.

- Have the experts describe each of the categories & sub-categories in terms of features *which the expert believes are appropriate*. Effectively the expert provides us with a set of classification rules for the domain. (This knowledge source will be considered to be a perspective which the expert holds on this domain.)
- All the REFINER systems require the domain expert to assign a category (a label) to each of the instances. We are continuing with this practise here as it gives us a further perspective on the set of cases / instances.
- Compare the expert's two perspectives on the domain; namely, the rules the expert has articulated for each of the categories versus the annotations he /she has associated with each of the cases.
- We have implemented a system, INSIGHT, which compares these two perspectives. So instead of using a machine learning algorithm as the core of the system we are, in this approach, using a system to check the consistency between the 2 expert-provided perspectives.²

As noted above, INSIGHT is a development of the REFINER family of systems, yet incorporates a somewhat different approach. Whereas the REFINER systems are able to infer descriptions of categories from a set of instances and to detect inconsistencies and suggest how they might be resolved, the INSIGHT system highlights discrepancies in two perspectives of an expert on a particular (classification) task, and brings these to the attention of the domain expert. In particular this realization of the checking tool, INSIGHT is able to handle annotated cases where the expert assigns each instance to one of the pre-designated set of categories. The second source of information is a set of rules which are able to classify each of the cases / instances. INSIGHT displays the results of such comparisons as a confusion matrix; an example of a confusion matrix for this domain is shown in Figure 2. The first row of the matrix consists of all the case which have been classified by the domain expert as "A"s whereas the cell (A, B) corresponds to cases which have been annotated by the expert as an "A" but have been classified by the rule set as a "B". Similarly the cases in the right hand cell in that row, cell (A, E), have been annotated by the expert as "A" but have been classified by the rules set as "E"s. Clearly all the diagonal cells [ie (A, A) (B, B), ... (E, E)] contain instances which have been classified identically by both the expert's annotation & by the rule-set.

² We shall see later that one of INSIGHT's modes does use machine learning techniques.

	Expected: A	Expected: B	Expected: C	Expected: D	Expected: E	Expected: (none)
Observed A	90 % 181 of 202	4 % 9 of 202	1 % 3 of 202	(none)	0 % 1 of 202	4 % 8 of 202
Observed B	0 % 4 of 1053	96 % 1014 of 1053	2 % 22 of 1053	0 % 4 of 1053	0 % 2 of 1053	1 % 7 of 1053
Observed C	(none)	4 % 22 of 533	87 % 462 of 533	6 % 34 of 533	0 % 1 of 533	3 % 14 of 533
Observed D	(none)	2 % 6 of 360	8 % 29 of 360	83 % 297 of 360	2 % 6 of 360	6 % 22 of 360
Observed E	(none)	4 % 20 of 540	2 % 11 of 540	12 % 67 of 540	78 % 423 of 540	4 % 19 of 540
Observed (none)	(none)	22 % 16 of 73	4 % 3 of 73	8 % 6 of 73	1 % 1 of 73	64 % 47 of 73

Figure 2: A confusion matrix

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

The Confusion Matrix (CM) seems to be a very intuitive way of presenting the results to experts; so far all the experts who have used it, have had no problem understanding it. Additionally it suggests a procedure for tackling the revision 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 (A, E) can be considered to be more surprising than those only one category away, say those in cell (A, B). Thus this distance measure suggests that the domain expert should be encouraged to consider discrepancies in the following order:

- (A, E) & (E, A); (Distance between categories of 4).
- (A, D), (B, E), (E, B) & (D, A); (Distance between categories of 3)
- (A, C), (B, D), (C, E), (E, C) (D, B) & (C, A); (Distance between categories of 2)
- (A, B) (B, C) (C, D) (D, E) (E, D) (D, C) (C, B) & (B, A); (Distance between categories of 1)

A further strategy which we suggested to the domain experts was for the first period to concentrate on removing the discrepancies from the *data-set* (incorrect annotations & 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 data set are localized, whereas a change to a rule could, in principle, effect *all* of the instances / cases.

A third strategy suggested was initially to refine each of the patient data sets *individually*, before attempting to refine the complete set of instances.

4.1 The Rule Interpreter

Essentially each rule consists of a set of one or more conjunctive conditions, and a single action which is to assign a particular instance to a category. 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 at the bottom of the list. In many situations the 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 has related conditions, then it is important to ensure they are appropriately ordered.

We have kept the format of the rules and the rule interpreter simple for a number of reasons: firstly, this meant the system could be implemented quickly; 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.)

4.2 Inferring Rules from Instances

INSIGHT has a mode which infers a rule when it is 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 instances. Our recent work with INSIGHT has made us aware of the need to select relevant descriptors from the inferred rule, in order to achieve effective distinctions between the categories. So even in this mode, we believe the process will require some involvement by the domain expert who will need to refine each rule by, for example, selecting descriptors from the set inferred by the Machine Learning algorithm.

This mode has still to be used by a domain expert with a demanding application.

5. Use / Evaluation of the INSIGHT system

Section 2 gives an overview of the evaluation to be undertaken; as mentioned in that section, the system's administrators provided us with a spreadsheet which contained the complete ICU stays for 10 patients. Each of these records was de-identified before this information was passed to us. Table 5 gives the code name for each of the patients and the number of recorded time points associated with each patient.

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 ex-

ample, 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 actual activities, nor all of the patient’s physiological parameters. Thus 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 this analysis (note that after the first 6 hours in the ICU, a complete set of “core” parameters is normally collected for the patient). It should be noted that some of the descriptors in this dataset (such as urine output and heart rate) were extrapolated to fill in certain missing values; the algorithm used to calculate these missing values was agreed with the clinicians.

Patient Code	696	705	707	708	720	728	733	738	751	782
Number of time points	129	576	475	40	188	281	396	110	493	73

Table 5: Patient codes and the number of records provided for each patient; there being in total 2761 patient records.

This section describes a two-stage study conducted with clinician-1 (sections 5.1 & 5.2), and a related 1-stage study undertaken with clinician-2 (section 5.3).

5.1 Review of Study with Clinician-1 (Phase-1)

Clinician-1 (MS) 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 his initial rule-set, however if the unclassifiable records are ignored that figure becomes 45.7% (258/564).

Further, at the end of the session (with just this one patient) the agreement was 97.0% (559 of 576) or 100% (556 of 556) if we ignore the effects of the (20) unclassifiable records. This session took about 5 hours, and was relatively slow as this was the first time INSIGHT had been used “in anger”, 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 such values. This and other functionalities have subsequently been added, so the 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, Mean Arterial Pressure (MAP), Noradrenaline and SpO₂.

In section 2.1, we outlined the nature of the knowledge available in this domain, and in section 4 we outlined the simple rule interpreter which we have implemented. Here we give some examples of the rules which have been implemented for several categories. For example, the rule associated with category A has the following form:

HR (normal-range) AND BLOOD-PRESSURE (normal-range) AND SPO₂ (normal-range) AND FIO₂ (normal-range) AND ADRENALINE (none) AND NORADRENALINE (none)

Note this is a *conjunctive* rule, and all the conditions have to be satisfied before a time point is classified as an “A”.

On the other hand, there are a number of disjunctive rules which represent each of the conditions which correspond to a patient being assigned to category “E”, namely:

HR (extremely low) OR HR (extremely high) OR MAP (extremely low) OR MAP (extremely high) OR ADRENALINE (extremely high) OR NORADRENALINE (extremely high)

Expert	Rule A	Rule B	Rule C	Rule D	Rule E
A		157 A→B* 2 A→C 3 rule edits 2 left as original annotation	11 A→B 7 A→C* 1 A→D	2 A→D* 2 rule edits	3 data edits ³
B	14 B→A*		1 rule edit 1 B→U 11 B→C*	2 B→C 1 B→D*	1 data edit ² 1 B→E*
C		12 C→B*		15 C→D* 1 C→E 2 rule edits 4 left as original annotation	1 C→E*
D		13 D→C 4 D→E 3 D→U	11 D→C* 1 D→E 2 D→U 1 rule edit		21 D→E* 1 rule edit 1 data edit ²
E					

TABLE 6: We have used the notation “nn A→B” to indicate that nn items which had been annotated initial 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.)

³ To remove impermissible values

The rules associated with categories B, C & D are also largely disjunctive, and tend to have values on a continuum from those associated with category “A” to those associated to category “E”, as section 2.1 suggests.

Clinician-1 (MS) followed roughly the refinement strategy suggested in section 4; note there are no E rows in this CM 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) & then (A, C) (A, B) (B, A) (D, E) (D, C) (B, C) (D, C) (B, D) (C, B) & (C, D). In the early stages of the analysis, very obvious inconsistencies were encountered & dealt with, and later it became often an issue of fine-tuning the rule-set and / or the data-set to achieve the classification which the expert wanted between two “adjacent” (1-distance apart) classifications. Table 6 gives a summary of the changes made to the “cells”.

Here we provide an overview of the typical decisions made by the domain expert:

- **Inadmissible Readings:** In cell (A, E), the expert considered that 3 of the values given in the data-set 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 he then had 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 feature 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 FiO_2 values of .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) etc); 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 300⁴ changes to annotations. Note some annotations might well have been changed several times: 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”.

⁴ 307 annotations were viewed, but 7 of these were left as the original annotation

In summary, this expert during the process of this refinement modified 52.1% (300/576) of his annotations. 7 changes (1.2%) were to reclassify an instance as unclassifiable (due to missing information, which in some cases the expert had overcome by “extrapolation” as discussed above); 274 changes (47.6%) were to adjust instances which were on the borderline between two of the A-E categories; and the remaining 25 (4.3%) were due to the expert overlooking a piece of information in the patient record which he accepted was important when it had been brought to his attention (by INSIGHT).

5.1.1 Rule Refinements

To date we have observed two significant types of rules / rule-sets 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 FiO_2 . Note that all the values returned for FiO_2 are effectively integers; also note that all the ranges for FiO_2 are continuous. Below we give the values for FiO_2 for a number of categories, both before and after refinement:

Category	Before refinement	After Refinement
C	0.55 – 0.69	0.55 – 0.69
D	0.70 – 0.84	0.70 – 0.83
E	0.85 – 1.00	0.84 – 1.00

Table 7: FiO_2 Values Before and After Refinement

Below we give an overall summary of the actions taken during this analysis:

Number of instances in the set	576
Number of instances / annotations viewed	307
Number of data values edited / removed	5
Number of annotations changed to unclassified	7
Number of annotations left as “inconsistencies”	7
Number of annotations changed to another A-E level (excluding “unclassified”)	46
Number of annotations changed to be consistent with the rules (excluding unclassified)	242
Number of changes to the rule-set	10

Table 8: Summary of Actions taken by Clinician-1 in Phase 1

5.2 Review of Study with Clinician-1 (Phase-2)

In this session, which lasted about 5 hours, we started with the rule-set which had been produced in this first session (when the expert had processed the data associ-

ated 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 (ie 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.

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 unclassifiable 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 unclassifiable instances had been removed, to 100.0% (2591 of 2591). The expert initially chose to view the same parameters as he did at the end of the first session, but part way through he added Dobutamine. The strategy followed by the expert for refining these instances was very similar to that given above.

Again we conclude this section by providing a similar summary to the one given in the previous section, see 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.

Number of instances in the set	2130
Number of instances / annotations viewed	225 ⁵
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 annotations changed to be consistent with the rules (excluding unclassified)	104
Number of changes to the rule-set	6

Table 9: Summary of Actions taken by Clinician-1 in Phase 2

⁵ This figure is approximate as there are several ways in which it could be calculated.

5.3 Review of Study with Clinician-2

In this session with clinician-2 (JK), which lasted about 2 hours, we started with the rule-set which had been produced in the second session by clinician-1 as the result of reviewing all 10 patients (see table 5). Further this clinician, clinician-2, had annotated three patient data-sets, 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, FiO₂, Gelofusin, Hartmanns, heart rate (HR), LiDCO Cl, MAP, Noradrenaline, PiCCO, Propofol, Sodium Chloride, SpO₂, temperature, urine output, and Vasopressin (18 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% agreement with the annotations created by this domain expert for patient 708, and by the end of this session the agreement had increased to 97.5%. These percentages are further improved, as one can see from Table 10 when the unclassifiable instances are removed. This table also gives results for patients 728 & 733 as well as for all three patients; in all cases results including & excluding unclassifiable instances, are reported. The percentage agreement, after data set & rule refinements, for all these data sets is remarkably high: being ~97.5% when unclassified cases are included and ~99% when they are not. Note too that initially 5 unclassifiable instances had been detected, after the refinement process this number rose to 11.

	P708 (before)	P708 (after)	P728 (before)	P728 (after)	P733 (before)	P733 (after)
All instances considered	40% 16/40	97.5% 39/40	10.7% 30/281	97.5% 274/281	8.1% 32/396	97.6% 387/396
Unclassifiable instances excluded	40% 16/40	100% 39/39	10.8% 30/278	99.6% 274/275	8.1% 32/394	98.7% 387/392

	All 3 patients (before)	All 3 patients (after)
All instances considered	10.7% 77/717	97.6% 700/717
Unclassifiable instances excluded	10.8% 77/712	99.6% 700/703

Table 10: Summary of Clinician-2's Refinement

5.4 COMPARISON between final data-sets & rule-sets for Clinician-1 & Clinician-2

The results in the diagonal cells of Table 11 are those for the individual clinicians and as such are reported at the end of sections 5.2 & 5.3 respectively. The figures in the off-diagonal cells give the agreements between the final rule-set & datasets of the 2 clinicians. As can be seen when unclassified instances are included in the analyses the results are 94.0% & 93.0% & when the unclassified items are removed from the analyses the agreement becomes: ~96% in both cases.

	Clinician-1's final <i>data-set</i>	Clinician-2's final <i>data-set</i>
Clinician-1's final <i>rule-set</i>	96.4% (2663 of 2761) 100.0%* (2591 of 2591)	94.0% (674 of 717) 95.9%* (674 of 703)
Clinician-2's final <i>rule-set</i>	93.0% (2567 of 2761) 96.3%* (2495 of 2591)	97.6% (700 of 717) 99.6%* (700 of 703)

Table 11: Comparison between final data-sets & rule-sets for Clinician-1 & Clinician-2. * These results correspond to analyses when the Unclassified instances are removed from the calculation.

These analyses suggest extremely high correlations between both the annotations & the rule-sets produced by these clinicians.

6. Contributions of this Work

- Produced a simple and useful tool to help experts appreciate how two perspectives on the same task is inconsistent and allows them to explore ways in which the two sources of knowledge can be made (more) consistent
- 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.
- Challenged the accepted wisdom of Cognitive Science (Expertise Studies) that a domain expert's "active" knowledge is more reliable than his "passive" knowledge
- Confirmed the need, when acquiring knowledge from domain experts, to determine whether a particular category has sub-categories & if so to get the expert to articulate them.
- Confirmed the need for sizable numbers of instances for each of the (sub)-categories when Machine Learning algorithms are used to infer associations.

- Confirmed the need to have a domain expert critically review any rules (knowledge) produced by an automated system. More particularly, INSIGHT has shown the benefits of experts being able to see their knowledge *applied* on a series of relevant tasks, and being able to comment on the outcome.

7. Further Work

The following are some of the activities planned:

Plan to evaluate the ICU scoring system across several ICUs & 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 / classification system. So far, we have applied INSIGHT to data from only 10 patients from a single ICU, and this information has been evaluated by just two domain experts. 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.

Excluding “unclassifiable” records from the analysis. Modify the rule-set such that all records which do not contain values for the core parameters will be excluded from the analysis. Before this can be implemented, a decision will have to be made about what constitutes this set of core parameters.

Use of INSIGHT with other domains. 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; INSIGHT should help with this process.

Extend INSIGHT so that it could be used to achieve consistency between more than 2 knowledge sources.

Use of INSIGHT’s mode to create rules from instances. We noted in section 4 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 not then need to create a set of rules which correspond to each of the categories. We need to test this hypothesis with a number of domains and with a range of experts.

Develop a variant of INSIGHT to apply 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.

References

- B. G. Buchanan and E. H. Shortliffe (Eds) (1984) *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*, Publ: San Francisco: Addison-Wesley.
- D. Corsar, & D Sleeman. (2007) *KBS Development Through Ontology Mapping and Ontology Driven Acquisition* Derek Sleeman, Ken Barker (ed), K-CAP '07: Proceedings of the 4th International Conference on Knowledge Capture (Whistler, BC, Canada): pages 23-30. ACM, New York, NY, USA.
- D Diaper (1989) *Knowledge Elicitation: principles, techniques & applications*. Publ: London: Ellis Horwood.
- K. A Ericsson, & HA Simon, (1993). *Protocol analysis; Verbal reports as data* (revised edition). Cambridge, MA: Bradfordbooks/MIT Press.
- O. Etzioni, M. Cafarella, D. Downey, S. Kok, A. Popescu, T. Shaked, S. Soderland, D. Weld, and A. Yates. (2005). Unsupervised named-entity extraction from the web: An experimental study. *Artificial Intelligence*, 165(1):91–134.
- P.E. Johnson (1983). What kind of expert should a system be? *Journal of Medicine and Philosophy*, 8, pp77-97.
- W. A Knaus, D. P Wagner, E. A Draper, J E Zimmerman , M. Bergner, P. G Bastos, C. A Sirio, D. J Murphy, T Lotring, A Damiano, et al. (1991). "The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults". *Chest* 100 (6): 1619–36.
- P Langley, & H. A Simon . (1995). Applications of machine learning and rule induction. *Communications of the ACM*, 38, November, 55–64.
- S. Marcus and J. McDermott. SALT: a Knowledge Acquisition Language for Propose-and-Revise Systems. *Artificial Intelligence*, 39(1):1_38, 1989.
- R. Michalski and R. Chilausky (1980) *Knowledge Acquisition by Encoding Expert Rules versus Computer Induction from Examples: A Case Study Involving Soybean Pathology*. *International Journal of Man-Machine Studies*, 12 (1) pp. 63-87
- L. Moss, D. Sleeman, J Kinsella, & M Sim. (2008) *ACHE: An Architecture for Clinical Hypothesis Examination*, Proceedings of 21st IEEE International Symposium on Computer-Based Medical Systems (CBMS 2008) (Jyvaskyla, Finland): pages 158-160

J. Pearl (1988). Probabilistic Reasoning in Intelligent Systems, Morgan Kaufmann.

J. R Quinlan (1986). Induction of decision trees. Machine Learning, 1. pp 81-106.

S Sharma & D Sleeman (1988). Refiner: A Case-based Differential Diagnosis Aide for Knowledge Acquisition and Knowledge Refinement. In Proceedings of EWSL-88 D Sleeman (Ed). Pitman: London, pp 201-210.

P Singh, T Lin, E Mueller, G Lim, T Perkins, & WL Zhu (2002) Open Mind Common Sense: Knowledge acquisition from the general public. In Proceedings of the First Intern. Conf. on Ontologies, Data bases, and applications of Semantics for large Scale Information Systems. Lecture Notes in Computer Science, Heidelberg: Springer-Verlag.

D Sleeman (1994). Towards a Technology and a Science of Machine Learning. AI Communications, 7(1), pp 29-38.

D. Sleeman, K. Barker, & D. Corsar (In press) Report on the Fourth International Conference on Knowledge Capture (K-CAP 2007), AI Magazine.

L. von Ahn, (2006). Games with a Purpose. IEEE Computer Magazine, pp 96-98.

Acknowledgements

- Sunil Sharma and Mark Winter implemented the earlier versions of REFINER.
- Andy Aiken & Laura Moss were both partially supported by EPSRC Studentships when they undertook this work.
- Consultants at the ICU at Glasgow Royal Infirmary for useful discussions & support on a range of related studies.

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 (ADR / NADR) 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 ADR / NADR and Oxygen.
- C.** Patient CVS system is effectively stable; probably on moderate dosages of ADR / NADR and Oxygen.
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 ADR / NADR/ 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 ADR / NADR 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