Resolving Differences of Opinion between Medical Experts: A Case Study with the IS-DELPHI System

Derek Sleeman[1], Kiril Kostadinov[1], Laura Moss[1,2] and Malcolm Sim[2]
[1]*Computing Science Department, The University of Aberdeen, AB24 3FX, U.K.*
[2]*Academic Unit of Anaesthesia, Pain and Critical Care, School of Medicine, University of Glasgow, G31 2ER, U.K.*

Keywords: Expertize Capture, Clinical Decision Support Systems, Inconsistencies, Negotiations, DELPHI-approach, Refinement, Cognitive Informatics, Medical Informatics, Pattern Recognition and Machine Learning.

Abstract : Knowledge intensive clinical systems, as well as machine learning algorithms, have become more widely used over the last decade or so. These systems often need access to sizable labelled datasets which could be more useful if their instances are accurately labelled / annotated. A variety of approaches, including statistical ones, have been used to label instances. In this paper, we discuss the use of domain experts, in this case clinicians, to perform this task. Here we recognize that even highly rated domain experts can have differences of opinion on certain instances; we discuss a system inspired by the Delphi approaches which helps experts resolve their differences of opinion on classification tasks. The focus of this paper is the IS-DELPHI tool which we have implemented to address the labelling issue; we report its use in a medical domain in a study involving 12 Intensive Care Unit clinicians. The several pairs of experts initially disagreed on the classification of 11 instances but as a result of using IS-DELPHI all those disagreements were resolved. From participant feedback (questionnaires), we have concluded that the medical experts understood the task and were comfortable with the functionality provided by IS-DELPHI. We plan to further enhance the system's capabilities and usability, and then use IS-DELPHI, which is a domain independent tool, in a number of further medical domains.

1 INTRODUCTION AND LITERATURE REVIEW

Knowledge intensive systems are pervasive across a wide range of industries (e.g. industrial, healthcare, automotive) and applications (e.g. expert systems, tutoring systems, adaptive interfaces). These systems often require domain knowledge which can then be exploited to address a range of tasks. Knowledge acquisition is the process of acquiring knowledge for subsequent use in knowledge intensive systems. One method to perform knowledge acquisition is through direct interaction with domain experts.

Experts' judgements are highly regarded, accurate, and can be characterized by: "employing more effective strategies than others; perceiving meaning in patterns that others do not notice; form rich mental models of situations to support sense making and anticipatory thinking; have extensive and highly organised domain knowledge; and are intrinsically motivated to work on hard problems that stretch their capabilities", (Klein et al, 2017).

However, there are a wide range of challenges faced in the elicitation of knowledge from experts in order to acquire accurate and useful knowledge, (Shadbolt & Smart, 2015). In this paper we focus on differences which may be found between experts' performance during a knowledge acquisition task. Although highly regarded, a domain expert may still display errors or inconsistencies in their performance. Cognitive biases (e.g. in deterministic and non-deterministic reasoning, (Tversky & Kahneman, 1974)) have been shown to occur in expert reasoning, (Fischhoff, 1989). Other conditions in which a domain expert's performance can be affected include: experts can 'gloss over' information i.e. miss surface detail and overlook features; context dependence within a domain i.e. without contextual-enabling information, experts may be more limited in their abilities; and inflexibility i.e. experts can have problems adapting, for example, to (significant) changes in basic domain concepts, (Chi, 2006). In dealing with uncertain scenarios, such as in the medical domain, people, including experts, often

Sleeman, D., Kostadinov, K., Moss, L. and Sim, M.
Resolving Differences of Opinion between Medical Experts: A Case Study with the IS-DELPHI System.
DOI: 10.5220/0008922000660076
In *Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2020) - Volume 5: HEALTHINF*, pages 66-76
ISBN: 978-989-758-398-8; ISSN: 2184-4305

make mistakes as they deviate from decision theory.[1] Although we have described how an expert's performance may be negatively affected, it is suggested that the observed effect is smaller than in non-experts, (Shields et al, 1987). Further, in some cases, biases can be considered as useful components of expertise; at least this is the perspective of the naturalistic decision-making movement, (Klein et al, 1986).

Several knowledge elicitation tools have been developed to help overcome some of the challenges faced during knowledge elicitation. Of particular interest to the work reported here are approaches which help experts detect and resolve differences when performing a task. One way to address the problem of differing opinions between experts is to perform a Delphi exercise during which perceived differences between experts are resolved through several rounds of discussions, (Murry & Hammons, 1995). Other methods to enable consensus with group decisions, include the Analytical Hierarchy Process, (Saaty, 2008) and the Nominal Group Technique, (Dunnette et al, 1963).

Software tools have been developed to help support such activities. Open Source ACH is a freely available implementation of the Analysis of Competing Hypotheses (ACH) method[2]. In the tool, a table (or matrix) is displayed with hypotheses presented in columns and evidence presented in rows. Each piece of evidence (a row) is then compared with each of the hypotheses (columns) and a level of consistency is noted. Aengenheyster et al, 2017 provide a survey of electronic "real-time delphi" tools with a focus on the following tools: Risk Assessment and Horizon Scanning (RAHS)[3], eDelfoi[4], Global Futures Intelligence System (GFIS)[5], and Surveylet[6]. The aim of these tools is largely to manage and conduct experiments using the Delphi process (e.g. by providing tools to create electronic questionnaires). For a detailed summary of other computerized tools which support group decision making see (Fjermestad & Hiltz, 1998). (On the other hand, we believe, IS-DELPHI is unique in creating visual displays of differences which in turn suggest changes that can lead to expert agreement.)

So far, the approaches and tools discussed largely support experts developing a consensus on whether they agree on propositions (e.g. "The use of antibiotics is set to increase over the next 5 years"). In a Delphi approach, this is achieved through repeated rounds until a pre-determined level of agreement has been reached. However, such group exercises often appear disjoint from the context of knowledge application, and further may not allow sufficient investigation of various opinions. Supporting electronic tools provide limited functionality for experts to *justify their opinion* (i.e. to offer insights into how or why they formed their opinion) or to *directly discuss points of disagreement* with other expert participants. Tools which support other types of knowledge elicitation, such as procedural knowledge or classification, are effectively lacking. For example, Open Source ACH provides collaborative features which help experts identify the basis of a disagreement, but it is limited to hypothesis generation and identification of supporting evidence (e.g. in intelligence gathering).

In this paper we describe a tool, IS-DELPHI, which is loosely based on the Delphi approach, but has been designed to support the development of an expert consensus on the *performance of classification tasks* rather than propositional statements. Further, to enhance the likelihood of achieving consensus during such problem-solving tasks, the facility for experts to *discuss* how they have classified an entity is included as an integral part of this tool.

The rest of the paper is organized as follows: section 2 summarizes the IS-DELPHI system which supports experts as they discuss differences with fellow experts; section 3 gives an overview of IS-DELPHI's functionality; section 4 reports a case study; section 5 discusses the results of that study; and section 6 discusses future work.

2 RESOLVING DIFFERENCES OF OPINION BETWEEN FELLOW EXPERTS: THE IS-DELPHI PROJECT

The focus of our group's approach has been to see if experts, given suitable tools, are themselves able to detect, and in some cases, resolve inconsistencies. In an earlier (medical) study, we asked clinicians to annotate a series of time-points indicating which class

[1] Namely, the principles of how to achieve optimum outcomes when reasoning or making decisions under uncertainty.

[2] http://competinghypotheses.org/

[3] www.rahs-bundeswehr.de

[4] www.edelfoi.fi

[5] www.themp.org

[6] https://calibrum.com/

each instance, in their view, belonged to. (These classes were intended to capture the ICU patient's overall physiological state.) In that study we mentioned the set of 5 classifications we thought should be used and asked each expert to annotate the set of instances independently. Once these annotations were available, we used a system we had implemented called INSIGHT (Sleeman et al, 2012) to compare a particular expert's annotations with those of a rule-set produced by a senior clinician from that group. This tool allowed the several experts to achieve agreement between their annotations and a ruleset in several ways; namely they could modify an instance's annotation, or they could revise the ruleset. At the heart of this system was a Confusion Matrix which showed the expert when the annotations produced by the single expert and their (evolving) ruleset agreed. Where there was agreement, that instance would appear in one of the diagonal cells; if there was disagreement then it would appear in an off-diagonal cell. Quite good agreements were obtained based on 3 experts, (Sleeman et al, 2012). To progress further we held a Delphi-style session with 3 medical experts and analysts who acted as enablers. The result of this session was a ruleset, approved by the 3-experts, which did provide better agreement than the individual rulesets, but we noted that inconsistencies (still) existed between the "common" ruleset and the annotations produced by the individual experts who had taken part in producing the common ruleset [7], (Sleeman et al, 2012). These experiences led us to consider holding sessions between **pairs** of clinicians where they had annotated a number of instances differently, to see if it might be possible for these differences to be resolved – or perhaps for individuals to clearly explain why they were **not** prepared to change their annotations. That is, we decided it would be useful to carry out a series of Delphi-style interactions between pairs of experts.

Subsequently, we have developed the IS-DELPHI web-based system which attempts to carry out programmatically, the negotiation stage discussed above when it is presented with 2 sets of annotations of a common set of instances produced earlier by pairs of experts. The investigators need to review the sets of annotations and decide which pairs of experts might find it useful to have Delphi-style dialogues /discussions. (Clearly, if a pair of experts produced *identical* annotations it would not be helpful for them to "meet".)

3 OVERVIEW OF THE IS-DELPH SYSTEM

The system provides facilities for the administrator(s) to set up studies between pairs of experts, and secondly provides facilities for pairs of experts to view their differences and attempt to resolve them. These functionalities are described in sections 3.1 and 3.2 respectively.

3.1 Facilities Provided for the Administration / Setting-up of Studies

There are 3 main functionalities provided here,

		Your Partner's Annotations					
		A	B	C	D	E	No Class
Your Annotations	A	None	None	None	None	None	None
	B	None	None	None	None	None	None
	C	None	None	None	None	None	None
	D	None	None	None	17.83% 23 of 129	None	None
	E	None	None	None	1.55% 2 of 129	80.62% 104 of 129	None
	No Class	None	None	None	None	None	None

Figure 1: This shows a typical Confusion Matrix displayed by IS_DELPHI which reflects the annotations made by 2 users.

[7] Subsequently we used Bayesian models to represent the annotation processes of each expert, and we discovered, in this formulism, there were distinct differences between the several expert models, and further that, occasionally, each expert failed to accurately apply their distinct model, (Rogers et al, 2013). (The later corresponding to what psychologists refer to as "slips".)

namely: displaying a list of active users, and the registration of new users; creating new projects and uploading the corresponding annotated datasets; and creating new studies between existing (registered) users and their already uploaded annotated datasets. Additional functionalities allow the administrator to archive the annotated CSV files produced by the experts, to download log files, and to terminate sessions etc. For more details see, (Sleeman & Kostadinov, 2019).

3.2 Facilities Provided to Enable Experts to Explore and Resolve Their Differences

There are 3 main groups of facilities available here, namely:

- Accessing IS-DELPHI[8], logging in to an existing account using your password.
- Studies: Pressing this TAB, lists all the studies which the user is currently involved in, shows the name of your partner(s), and provides a 'view' button which allows you to see the state of the Confusion Matrix of each active study. Figure 1 shows a screenshot of a typical Confusion Matrix.
- Ability to revise your annotations and to leave associated comments. This is a significant topic and is described in some detail in the rest of this section.

The Confusion Matrix in Figure 1 is quite sparse as the 2 experts have only used "D" and "E" annotations. The cells on the diagonal indicate agreement between the 2 sets of annotations, whereas off-diagonal cells show disagreements. (Note, the active user's annotations are indexed through the *rows* of the table, while the other expert's annotations are indexed though the *columns*.)

If a user clicks on any cell, the system opens another window at the top of the screen which displays all the time-points in that cell. Figure 2 shows the cell where User 1 (the active user) has annotated both instances as Es – and the other expert has annotated them as Ds.
User 1 (the active user) is then able to change his annotation if he wishes (using the drop-down menu shown in Figure 2). In fact the user is able to make as many changes as he wishes to this cell. In general, as there are several cells where the experts disagree, the active user can change the annotations associated with each of these. In the initial implementation, we only allowed a user to add a comment to a cell *after* an annotation had been changed. Once the user has reported all the annotations and comments they wish to make, then they can press the 'submit changes' button which then transfers control of the study to their partner. (As only one partner at a time has the ability to modify annotations), The system then sends the partner an email telling them that they can now access the system and make changes to the annotations if they wish. In fact before the system gives control to the other partner it asks the currently active user one further question – namely "Have you made all the changes to your annotations, which you think, are necessary?" As mentioned above, the ability to make changes to annotations and add comments then passes to the partner, who can then make such changes if they wish or they too can indicate, by pressing the "NO FURTHER CHANGES" button, that they are satisfied that no further changes are necessary. Note the system only closes a study if both users press the "NO FURTHER CHANGES" button in 2 *consecutive* interactive sessions. Again, for more details see (Sleeman & Kostadinov, 2019).

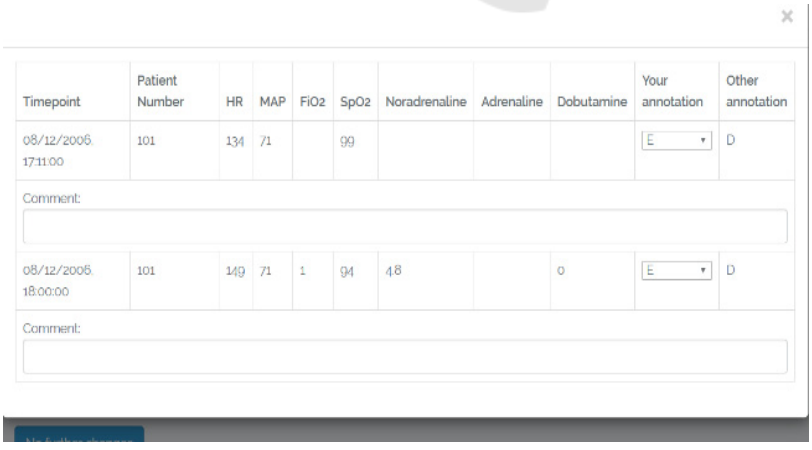

Figure 2: This shows details of an off-diagonal cell – where the 2 experts disagree on the annotation.

[8] http://intelsys-abdn.org.uk

4 CASE STUDY

As noted earlier, we have developed a scoring system which captures the overall state of a patient in the Intensive Care Unit (ICU PSS), and can be used, unlike APACHE, at any point in a patient's ICU stay, (Sleeman et al, 2012). The score uses commonly available physiological and pharmacological data and aims to take into account the amount of support the patient is receiving. That is, a patient with a "normal blood pressure" but receiving large quantities of inotropes, would not be classified the same as a patient with a "normal blood pressure" but no pharmacological support. By means of a sophisticated rule base the scoring system summarises the overall state of the patient from A to E. Where the **"A"** level represents a relatively well patient on minimal, or no physiological and / or pharmacological support, who may be ready for transfer to a ward or a high dependency unit. On the other hand, **"E"** represents a patient who is extremely sick requiring considerable physiological and / or pharmacological support. **"C"** represents a patient who is somewhere in the middle i.e. moderately unwell and receiving some physiological and / or pharmacological support.

The case study was run as 2 parts: firstly, the pilot study was run with 4 participants[9]. Results from the questionnaires and observations from the analysts (DS and KK) were used to enhance the system and some aspects of the documentation, and then a "full" study was run with 12 ICU clinicians at the Consultant or Senior Registrar levels. The overall structure of both studies was the same, namely:

- Each participant was asked to annotate independently an identical set of 60 time-points for ICU patients (these had the same form as the clinicians see regularly in the ICU) as one of 5 categories. The participants added their annotations to a spreadsheet. (So, this information was captured on-line.)
- The analysts then reviewed these responses, and selected pairs of experts who replied differently to a number of time-points. (In the pilot this selection was done "manually"; for the "Full" study we used software to help make these selections.)
- Use of the IS-DELPHI system: the system administrator then registered the various experts and loaded their annotated files to the system. Further, they set up a series of "studies" each of which involved a pair of experts identified in the previous step.

Although, IS-DELPHI has been designed as a web-based system, in the pilot study, the participants were asked to come to a central laboratory in the Queen Elizabeth University Hospital complex (Glasgow) so that the investigators could better monitor the interactions between the participants.

4.1 Results of the Pilot Study

The comments below are based on feedback contained in the questionnaires returned and more specifically by a detailed report provided by the medical author of this paper (MS) who was also a participant in the pilot study:

- Various enhancements to the user interface were suggested, so the expert can clearly indicate when they had completed revising the annotations associated with a particular cell.
- Comment facility: As noted above (section 3.2) the initial implementation of the system only allowed a user to add a comment once an annotation had been changed. This expert argued that it should be possible to add comments about a time-point before a change is made. Indeed, he argued that this facility should be much more flexible – including allowing an expert to view, in a separate window, a complete interaction (annotations and comments) between the 2 experts – where each interaction is time-stamped.
- A significant discovery was that some browsers were not communicating effectively with IS-DELPHI which resulted in some annotation changes and comments not being recorded by the system.

We determined which browsers were causing problems, and then blacklisted these in the User manual. Secondly, the various system changes suggested were considered by the analysts and the majority were implemented before the full study was undertaken.

4.2 The Full Study

As noted in section 4, 12 ICU specialists were involved in the full study which took essentially the same form as that noted above for the pilot project (namely, annotation of a common set of instances, selection of

[9] 2 ICU specialists and 1 trainee anaesthetist took part in this study, and to complete a pair one of the analysts (DS) also participated. DS used the definitions defined by experts in the earlier ICU PSS study with INSIGHT, (Sleeman et al, 2012) as his "crib".

pairs of experts to resolve their differences, and use of IS-DELPHI by these pairs.)

4.2.1 Selecting Pairs of Experts Who Might Usefully Use the IS-DELPHI System to (Attempt to) Resolve Their Differences

As noted above, this study was run with 12 experts who were asked to annotate an identical set of 60 instances. To select appropriate pairs, the Python-based COMPARE program firstly creates a table which shows the number of agreements / disagreements between each of the pairs i.e. (P1 P2), (P1 P3) … (P1 P12), (P2 P3), (P2 P4)… (P2 P12) .. (P11 P12). (Matching Criterion-1: 2 Annotations are considered *different* if they are ***1 or more*** ICU PSS categories apart.) Note too that P3 did not return any annotations which in turn means that only 5 groups can be formed. This information is shown in Table 1 where we report the number of instances on which the pairs of annotations are different with respect to the current matching criterion. For example (P1, P2) reports 34 disagreements which means there are 26 (60 – 34) **agreements**. The average number of differences shown in this table is 31.69 with a SD of 9.35; with those figures becoming 27.80 and 4.60 respectively when P11, who seems to be an "outlier", is removed from the calculation.

The next step was to select pairs of experts who would work together to resolve their differences. Here, the algorithm chose 5 pairs which gave overall the *smallest* number of instances to be considered, provided that each pair contained at least the minimum number of instances specified (2 in this study) where this value is passed to the COMPARE algorithm as the *threshold* parameter. And of course each instance selected, complies with the current Semantic Matching criterion. (Clearly, we are not interested in selecting pairs, which have no differences, as in such cases the experts would have nothing to discuss in IS-DELPHI.) Given that there are a relatively small number of participants in this study, it would have been possible to select these pairs "manually". However, as the process is somewhat time-consuming and error-prone, we decided to add further functionality to the COMPARE program to make these selections. This selection process is based on a maximum weighted matching algorithm, (Galil, 1986) which attempts to minimize the number of items in all the chosen pairs, with the added constraint, noted above, that each pair **must** contain at least 2 instances. The pairs chosen by COMPARE for this study are highlighted in Table-1.

The number of instances selected, by COMPARE, were considered too large for the expert-pairs to process (using IS-DELPHI), and so we decided to investigate the number of instances where the semantic *differences* are 2 or more ICU PSS categories apart. *(Matching criterion-2).* In this scenario, (A C) (A D) (A E) (B D) (B E) (C E) and the reversed forms would be reported but (A B) (B A) (B C) (C B) etc would not be. When this criterion was applied the number of disagreements between the pairs of experts, reported by the COMPARE system, reduced quite considerably as shown in Table 2. The average number of differences shown in this table is 7.00 with a SD of 9.59; with those figures becoming 2.76 and 2.99 respectively when P11, who seems to be an "outlier", is removed from the calculation. Again the COMPARE system selected pairs of experts who have 2 or more time-points. For example, in Table-2 the pair of experts (P1, P2) are

Table 1*:* Number of Disagreements (Using current criteria). The selected pairs are highlighted in YELLOW; note P11 is not assigned to a pair because there are an odd number of participants – also this person does seem to be an outlier. Pn indicates the nth Participant in this study.

	P1	**P2**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**	**P11**	**P12**
P1	-	34	24	28	23	22	31	29	32	49	23
P2		-	25	20	31	37	27	32	27	52	36
P4			-	21	26	30	25	28	25	50	22
P5				-	33	34	26	23	34	49	34
P6					-	26	28	34	27	47	19
P7						-	26	27	30	50	28
P8							-	19	32	49	26
P9								-	31	52	30
P10									-	50	26
P11										-	44
P12											-

shown as having 3 instances which have a "semantic difference" of at least 2 classes between their annotations. (It is not uncommon, in this domain, for annotations produced by 2 experts to differ by **one** category.) The pairs highlighted in Table-2 were selected by the algorithm, namely: (P1 P6) (P2 P5), (P4 P10), (P7 P9) & (P8 P12); it is only possible to form 5 groups; P11 was excluded by the criteria we have used for pair selection. Moreover, looking at the "raw" data it would appear that P11 is, anyway, somewhat of an outlier when compared to the rest of this group. (In future studies, once the general approach has been shown to be workable, we might decide to include "outliers" like P11 to see if this approach would work with pairs of experts which have a substantial number of differences. Also see section 6.)

The COMPARE system has 3 parameters, namely: whether unlabelled classes are to be handled, the minimum "semantic" distance between 2 annotations for them to be considered ***not*** to match, and the minimum number of instances to be considered by each pair of experts. Note, the algorithm reports the number of mismatches, *not* matches, found.

4.2.2 Use of the IS-DELPHI System

The instructions to the participants in this study recommended that they consider time-points where the annotations provided by the experts, are 2 ICU PSS classes different, as differences of just one class apart are not uncommon. (The experts could, of course, choose to consider **any** of the time-points contained in the confusion matrix.)

The various changes made and comments recorded by each of the experts participating in a particular study are recorded by the system in a log file which is unique to each study. Each interaction is time-stamped, and the system also effectively records when the "active" user changes and when a study is closed. These log files are discussed in some detail in section 4.3.

4.3 Results of the Full Study

The critical issue to be decided was whether IS-DELPHI was able to help these experts resolve their differences. Table 3 indicates strongly that this is indeed the case, and reports that from the 5 pairs there were 11 time-points where the 2 experts disagreed initially by 2 ICU PSS classes in their annotations, and after using the system 7 of these were **fully** resolved. Moreover, in the remaining 4 instances the differences were reduced to just 1 ICU PSS class which we had suggested in the documentation given to the study participants, in this domain, could be considered to be effectively equivalent. Further, there were **no** instances in which neither expert made a change.

We have also carried out a more fine-grained analysis which investigates how these changes were achieved – i.e. was an agreement achieved by just a single expert making a 2-step change or did both experts make 1-step changes? Of the 7 time-points where full agreement was achieved we note that 3 were achieved by a single expert making a "big" change, and the other 4 were achieved by both experts each making a smaller change. And of course, in the case of the 4 time-points where the final difference was 1 ICU PSS class, just one of the experts made a smaller change of 1 semantic class.

Table 2: The selected pairs are highlighted in YELLOW; note P11 is not assigned to a pair (because there are an odd number of participants – also this person does seem to be an outlier). (NB Matching criterion-2 has been used here, and each selected pair must have at least 2 instances.) Pn indicates the nth Participant in this study.

	P1	**P2**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**	**P11**	**P12**
P1	-	3	0	1	2	0	1	2	6	26	1
P2		-	0	2	3	8	4	10	11	30	1
P4			-	0	1	0	1	1	3	27	0
P5				-	3	3	1	1	9	22	1
P6					-	1	1	4	4	23	0
P7						-	2	2	4	31	0
P8							-	0	10	24	2
P9								-	8	23	3
P10									-	31	4
P11										-	24
P12											-

4.3.1 Analysis of Time-points Which Only Differ by One Semantic (ICU PSS) Class

Below we give a summary of the number of such instances which were presented to each pair and the number which each pair processed:

- F1: There were 21 instances of this type, of which 8 annotations were amended by just 1 expert.
- F2: There were 18 instances of this type, none of which were amended by either expert.
- F3: There were 22 instances of this type, of which 10 annotations were amended by the first expert and 11 by the other expert. The annotation of the remaining instance was not changed but one expert left several comments about it.
- F4: There were 25 instances of this type, of which 11 annotations were amended by one expert and none by the other expert.
- F5: There were 24 instances of this type, of which 4 annotations were amended by just one expert who also left a comment explaining why he was not changing a further annotation. The second expert made no changes or comments.

4.3.2 Analysis of the Actual Interactions Which Occurred between Experts When using IS-DELPHI

As noted above (section 3.2) the first implementation of IS-DELPHI used in the Pilot Study, only allowed experts to make a comment *after* they had changed an annotation. The feedback from the Pilot Study was that this is far too restrictive, so we amended the system to allow either comments, or a change of annotation, or both to be made at each stage of the interaction. This seems to have been a very valuable enhancement to the system, as the types of interactions we have seen in the full study are certainly much more flexible than the first system supported, and can be summarized as the expert user:

- Only modifies the actual annotation(s) - no comments provided
- Makes change to an annotation and adds a comment in the same interaction (in either order).
- Comments precede changes (these are given in different interactions)
- Changes precede comments (these are also given in different interactions)

There are many examples of these different "styles" of interactions in the log files collected from the 5 pairs of experts during this study. Figure 3 gives an excerpt of the interactions between experts P02 and P05 which includes several of these types of interactions – particularly the one which has changes of annotation and comments occurring in the same interaction. In this study, the 2 experts in fact resolved differences in their initial annotations for both TP-45 and TP-59. However to make the interaction easier to follow we have only included interactions which discuss TP-45; also we edited the dialogue a little to remove several typos etc. It is also important to know that P02's initial annotation of this time-point was an "E" and P05's was a "C". What's interesting about the interaction is that both experts summarize the "case" as they see it, allude to the reasons why they gave the initial annotations they did, and then discuss issues raised by their partner which is making them reconsider their initial annotations. With this instance (TP-45), the final interactions of both experts was to propose a compromise classification of "D", hence achieving *full* agreement on this time-point.

4.3.3 Feedback from the Participants in the Full Study

Stage 1: The questionnaire posed the following questions:

1. I understood the overall approach of the IS-DELPHI project. (Rating scale where 1 is very unclear, and 10 is absolutely clear):

Table 3: Across the 5 pairs there were 11 instances where the initial annotations of the 2 experts differed by ICU PSS 2 classes. The results show the agreement after each pair has used the IS_DELPHI system.

	Pairs				
	F1	**F2**	**F3**	**F4**	**F5**
First Time-Point	Full agreement	Full agreement	Full agreement	1 class different	1 class different
Second Time-Point	1 class different	Full agreement	Full agreement	Full agreement	1 class different
Third Time-Point	-	-	Full agreement	-	-

2. I understood what I was required to do in Phase 1 of the IS-DELPHI project. (Rating scale where 1 is very unclear, and 10 is absolutely clear):

Where the results are: Q1, Mean 9.00, standard deviation (SD) 1.26; Q2, Mean 9.18, SD 1.33. NB There were 11 participants; one expert completed neither the annotations nor the questionnaire.

The participants were also asked to make suggestions about the study. The following comments were received:

- 2 comments about details of units for the vasopressors
- A question about FiO2 and whether it applied to ventilated / unventilated patients.
- "*Very interesting and thought provoking study*."

Stage 2: This questionnaire posed the following questions:

1. I understood the overall approach of the IS-DELPHI project. (Rating scale where 1 is very unclear, and 10 is absolutely clear):
2. I understood what I was required to do in Phase 2 of the IS-DELPHI project. (Rating scale where 1 is very unclear, and 10 is absolutely clear):
3. The section of the user manual for domain experts was clear. (Rating scale where 1 is very unclear, and 10 is absolutely clear):

Where the results were: Q1, Mean 8.40, SD 0.84; Q2, Mean 8.00, SD 0.94; Q3, Mean: 7.10, SD 1.10. NB Only 10 people (5 pairs) took part in this phase.

The participants were also asked to make suggestions about the study. Only 2 comments were received:

- "*..I think the screen shots need to be clearer. There is a lot of explanation which is good but is a lot to take in and is quicker/easier to understand by practicing live. Perhaps earlier reference to opening up a version to look at and play with could help*."
- "*It was clear enough. Didn't need to refer too carefully to the manual, though my partner did*."

[2019-05-30 15:03:51] Expert **P02** left a comment on timepoint **TP-45** [PN-7465: 2017-10-07 15:00:00][1]: "Low oxygen saturations on maximum FiO2, quite fast heart rate with low/normal BP"

[2019-05-30 15:20:02] Expert **P05** left comment on timepoint **TP-45:** "Agreed, but not obviously needing CVS support. The picture I have is of a struggling respiratory failure patient requiring intubation. MAP of 58 might hint at impending CVS compromise but in absence of other data (e.g. CRT) hard to assume"

[2019-05-30 15:28:11] Expert **P02** left comment on timepoint **TP-45:** "I agree that is probably the most likely picture though this could be the situation in an intubated patient. Given the absence of inotrope/vasopressor requirement happy to change to D."

[2019-05-30 15:30:52] Expert **P05** left comment on timepoint **TP-45:** "Agree. D seems fair given both our 'angles'."

[2019-05-30 15:31:38] Expert **P02** changed evaluation on timepoint **TP-45** from E to D. No comment was provided.

[2019-05-30 15:35:12] Expert **P05** Changed evaluation on timepoint **TP-45** from C to D. Comment: "Okay, agreed to go to D."

. . . . . .

[2019-05-30 20:59:07] Expert **P02** indicated that they have no further changes.

[2019-05-31 14:20:17] Expert **P05** indicated that they have no further changes. The study ends.

Figure 3: This gives an excerpt of an interaction, captured by IS_DELPHI, between experts P02 and P05. This dialogue has been slightly revised with typos etc removed. In the complete session they also discussed a further time-point on which they initially disagreed, but for ease of presentation those comments have been removed – just leaving a discussion of a single time-point (TP-45).

5 DISCUSSION

In this paper we have described and illustrated the use of the IS-DELPHI tool, which has been designed to support the development of consensus between pairs of experts when solving common *classification* tasks. We believe that the provision of a facility which allows experts to discuss their several approaches to common tasks, has greatly contributed to the tool's effectiveness.

IS-DELPHI has been evaluated in a study with medical domain experts. The results of the full study show that the 5 pairs of clinicians were able to make considerable progress in agreeing on (common) annotations. In 7 out of the 11 instances, they achieved full agreement on the final annotations, and in 4 further instances reduced the differences between their annotations from 2 ICU PSS classes different to just 1 (which, we'd noted, in this domain, is commonly accepted as a match.) Our analysis of the interactions / dialogues between the several experts suggests that IS-DELPHI has elicited useful dialogues between the pairs of experts which has facilitated them reaching agreements on their revised annotations, (see Figure 3).

Further, we note that 4 out of the 5 pairs of experts also "processed" time-points which were only 1 ICU PSS class apart, something which was **not** specified in the study instructions. For instance, pair F3 instead of revising just the annotations associated with the 3 time-points whose ICU PSS classes were 2 semantic classes apart, actually made changes to 24 time-points (21 of which were just 1 semantic class apart). This suggests that many of the participants, understood the tasks well, and secondly, were quite comfortable with the functionality provided by IS-DELPHI.

Additionally, questionnaires from both phases suggest that the majority of the participants understood the task to be performed, and were happy with the instructions for both phases of the study and with the user manual provided. We have noticed that the lowest mean score and the greatest spread was for question 3 about the effectiveness of the manual. From studying the logs the analysts determined that some users were unclear how to terminate sessions.

In summary, we suggest that the IS-DELPHI system has a useful role to play in the important Knowledge Acquisition process; we have shown in this study that it enables clinical experts to discuss and often resolve their differences, so ensuring that the resulting medical knowledge is more consistent.

6 FURTHER WORK

Following this successful case study, we are now planning a number of additional activities including:

a) Other studies: a larger study based on ICU datasets, further medical domains and in other disciplines such as Management.
b) Adding a feature to process the experts comments / justifications with Natural Language and Argumentation Techniques.
c) Store more complex multi-stage comments / justifications.
d) Extend the criteria which the system uses to select pairs of experts, to include pairs with a larger number of differences.
e) IS_DELPHI has focused on tasks where the experts **classify** the various instances. We plan to investigate whether the system could be extended to also address further problem-solving approaches, e.g., planning.
f) Systems developments (archive log files, the export and import of the expert-refined CSV files, and address the several system enhancements mentioned in section 5.)
g) Create a video showing how to use the system, which will either supplement or replace the current user manual.

ACKNOWLEDGMENTS

We are grateful to the following colleagues from the QEUH (Glasgow) for taking part as experts in this study (2 in the pilot and the remainder in the full study): Dr. Russell Allan, Dr. Richard Appleton, Dr Aporva Ballal, Dr. Matthew Baynham, Dr. Euan Black, Dr. Andrew Cadamy, Dr. Alan Davidson, Dr Robert Hart, Dr. Louise Hartley, Dr. Andy Mackay, Dr. Johnny Millar, Dr. Peter Stenhouse, and Dr. Laura Strachan.

IS-DELPHI was implemented by Kiril Kostadinov with financial support from the University of Aberdeen Development Trust. Helpful discussions with Professor Wamberto Vasconcelos (Computing Science Dept., The University, Aberdeen) and Dr. Martin Shaw (Clinical Physics Department, Glasgow Royal Infirmary).

REFERENCES

Aengenheyster, S., Kerstin Cuhls, Lars Gerhold, Maria Heiskanen-Schüttler, Jana Huck, Monika Muszynska,

Real-Time Delphi in practice — A comparative analysis of existing software-based tools, *Technological Forecasting and Social Change*, Volume 118, 2017, Pages 15-27.

Chi, M.T.H. Two Approaches to the Study of Experts' Characteristics. In: *The Cambridge Handbook of Expertise and Expert Performance. Ericsson*, K.A., Charness, N., Feltovich, P.J., Hoffman, R. (Eds) Cambridge University Press. 2006.

Dunnette, M.D., Campbell, J.D., and Jaastad, K. The effect of group participation on brainstorming effectiveness for two industrial samples. *Journal of Applied Psychology*, XLVII pg 30-37. 1963.

Fischhoff, B. "Eliciting Knowledge for Analytical Representation," *IEEE Trans. Systems, Man, and Cybernetics*, vol. 19, no. 3, 1989, pp. 448–461.

Fjermestad, J., Hiltz, S.R. An assessment of group support systems experimental research: methodology and results. *Journal of Management Information Systems*. Vol 15. Issue 3 (December 1998).

Galil, Z., Efficient algorithms for finding maximum matching in graphs, *ACM Computing Surveys*, 1986.

Klein, G., R. Calderwood, and A. Clinton-Cirocco. Rapid Decision Making on the Fire Ground, *Proc. Human Factors and Ergonomics Soc. Ann. Meeting*, vol. 30, no. 6, 1986, pp. 576–580.

Klein, G., Shneiderman, B., Hoffman, R.R. and Ford, K.M. (2017, November/December). Why expertise matters: A response to the challenges. *IEEE: Intelligent Systems*, pp. 67-73.

Murry, J., J. Hammons, "Delphi a versatile methodology for conducting qualitative research", *Rev. Higher Educ.*, vol. 18, no. 4, pp. 423-436, 1995.

Rogers, S., Derek Sleeman, & John Kinsella. Investigating the disagreement between clinicians' ratings of patients in ICUs. *IEEE Journal of Biomedical and Health Informatics* 17.4 (2013): 843-852.

Saaty, T.L. Decision making with the analytic hierarchy process. *International Journal of Services Sciences*, Vol 1, No 1. 2008.

Shadbolt, N. R., & Smart, P. R. (2015) Knowledge Elicitation. In J. R. Wilson & S. Sharples (Eds.), Evaluation of Human Work (4th ed.). CRC Press, Boca Raton, Florida, USA.

Shields, MD., Solomon, I., and Waller, W.S. "Effects of Alternative Sample Space Representations on the Accuracy of Auditors' Uncertainty Judgments," *Accounting, Organizations, and Society*, vol. 12, no. 4, 1987, pp. 375–385.

Sleeman, D., & Kiril Kostadinov (2019). IS-DELPHI User Guide, Technical report, Department of Computing Science, The University of Aberdeen.

Sleeman, D**.,** Moss, L.**,** Aitken, A., Hughes, M., Sim, M., Kinsella, J. Detecting and Resolving Inconsistencies between Domain Experts' Different Perspectives on (Classification) Tasks. *Artificial Intelligence in Medicine*, 2012 Jun;55(2):71-86.

Tversky, A., & D. Kahneman, D. Judgement under Uncertainty: Heuristics and Biases. *Science*, vol. 185, Sept. 1974, pp. 1124–1131.