Experts Assess Patterns produced by a Temporal Discovery Workbench[1]

Derek Sleeman [1, 2] Sam Cauvin [1]
[1]Department of Computing Science,
University of Aberdeen,
Aberdeen, AB24 3UE
Email: {d.sleeman, s.cauvin} @abdn.ac.uk

Laura Moss [1, 2] John Kinsella[2]
[2]School of Medicine,
University of Glasgow,
Glasgow, G31 2ER
Email: {laura.moss, john.kinsella} @glasgow.ac.uk

ABSTRACT
An earlier study asked two experts to discuss conditions under which Myocardial Damage (MD) can occur in ICU patients; these experts were shown temporal records where some sequences resulted in MD and others which did not. The resulting model[2] was quite complex as it contained temporal constraints as well as the usual conjunctive and disjunctive terms. This was a classical KC Study. We have since implemented a Temporal Discovery Workbench (TDWB) to process the same temporal datasets to see if TDWB can discover simpler patterns to explain the same datasets. Subsequently, we have shown that the sets of patterns produced by TDWB generally have better "coverage", than those produced by the original model. We then investigated whether some of the TDWB-created patterns might not be clinically acceptable. Recently we ran a pilot study in which we asked a single clinician to evaluate the patterns produced by TDWB, and to say whether they were acceptable, and why. This further information has now been implemented in TDWB; the resulting set of filtered patterns still has better coverage than the initial set of "manual" patterns.

Keywords: *Modelling of Expertize, Temporal Datasets, Event Prediction, Workbench, Intensive Care Unit, Myocardial Damage.*

1. INTRODUCTION

Earlier, we reported a study where we asked experts to discuss conditions under which Myocardial Damage can occur in ICU patients and then compared the results of their model with a test dataset, [Sleeman et al, 2011]. Here's a summary from that report: "Myocardial damage is known to occur relatively frequently, and although it is not often fatal it results in the patient staying in the ICU for significantly longer. Thus it is important for clinicians to detect these events. Confirmation of myocardial damage is by a biomarker (troponin), but these tests are only done at fixed time-points. Consequently it is desirable for doctors, and support systems, to detect myocardial damage from the standard descriptors collected for ICU patients. We have undertaken a study with several ICU consultants to determine the conditions which generally precede a myocardial-damaging event. In fact, these knowledge acquisition sessions produced a complex model which we have realized as 2 modules. Subsequently, we compared this model's predictions against the original datasets; the model when run against the *test* dataset resulted in a high True Positive (TP) rate (75.8%)." [Sleeman et al, 2011].

This was a very encouraging result. However the model articulated by the experts as indicated above was relatively complex. The following is a slightly simplified summary of the conditions under which the experts believe Myocardial Damage (MD) occurs:

- **MD** is confirmed when a cardiovascular derangement (CVD) sequence is followed by a raised troponin value within [1-72] time-periods.
- A **CVD sequence** is said to occur when CVD (cardio-vascular derangement) events occur in at least 3 out of 5 *adjacent* time-periods
- A **CVD event** is said to be either:

  - A very extreme value for any of the following patient descriptors: SpO_2 (Oxygen Concentration in the patient's blood), HR (Heart Rate) or MAP (Mean Arterial Pressure i.e., the patient's Blood Pressure).[3] Note there are 5 possibilities as HR and MAP can have both extremely low and extremely high values.

  - A combination of 2 of the above descriptors with extreme values (Giving 8 combinations)

[1] This Workshop paper is based on Sleeman et al (2015); but does report a different analysis and outlines a series of studies in which domain experts evaluate the patterns produced by TDWB.

[2] In this project we refer to a model as a coherent set of patterns.

[3] For a detailed discussion of the descriptors recorded regularly for each ICU patient, and a scale used to describe the patient's status, see [Sleeman et al, 2009].

- A combination of 3 of the above descriptors with considerably abnormal values (but less severe than "extreme").

- Extremely high levels of FiO2 (inspired oxygen) or a rapid *increase* in FiO2 between several time-points.

As a result of this study we decided to develop the Temporal Discovery Workbench (TDWB) to see whether given background information about the domain, and the same temporal sequences as the experts analyzed, firstly the TDWB would be able to reproduce the (complex) model articulated by the experts and secondly whether it would be able to suggest some alternative, possibly simpler, models / patterns. We are addressing the general scenario in which an unusual event, E, happens at time-point, T, and we aim to predict this event by analyzing trends and absolute values in the several descriptors recorded in the time-period *prior* to E. To help this analysis, it is likely we will also have datasets involving the same descriptors in which the event, E, does **not** occur.

Advantages of Workbenches: As mentioned we have decided to implement a Workbench (Temporal Discovery Workbench – TDWB) as we believe this provides a great deal of flexibility. Specifically although we have a clear idea of the project's overall objectives we do not know in advance the range of applications we might encounter and thus we do not know the detailed nature of the analyses which domain experts might wish to carry out on their datasets. Workbenches (WBs) generally present their user with options at each stage in the analysis and allow the analyst (sometimes with guidance) to decide the data display mode or analysis package to be used and with what descriptors. It is essential that WBs provide user-friendly interfaces, and they are modular in construction, so that functionality not envisaged at the initial design can be subsequently added if needed. (In later sections we discuss the outline implementation of the TDWB.)

Overview of the Paper

Section 2 gives brief literature reviews of the analysis of temporal datasets and the Apriori algorithm. Section 3 outlines the functionality of the TDWB (Temporal Discovery Workbench). Section 4 reports the results of analyzing the Glasgow MD patient dataset with TDWB, and discusses a study when a domain expert provides feedback on the patterns generated. Section 5 discusses further work.

2. Literature Review

Temporal datasets are now regularly collected by many companies and institutions and there has been considerable interest in analyzing these datasets for example to detect inconsistencies, trends, recurrent patterns etc. Combi et al (2010) gives a good overview of temporal Information Systems in Medicine.

An important development in data mining has been the ability to establish that a descriptor is associated with one or more other domain descriptors; Agrawal & Skikant (1994) developed the very efficient Apriori algorithm to detect such patterns. Laxman & Sastry (2006) have subsequently developed this approach so that it is able to detect patterns in temporal datasets. TDWB attempts to infer association patterns between the descriptors in the (temporal) domain it is analyzing. The general form of the patterns/rules which it infers is:

IF A@T+0[4] and B@T+1 occur THEN expect E[T+2, T+50].

The Apriori algorithm – and more particularly the temporal extension are central to TDWB, this approach is outlined in more detail in Sleeman et al (2015).

3. Overview of the Temporal Discovery Workbench (TDWB)

Temporal datasets are presented to TDWB as CSV files, and must contain a column called Time-point (containing data of the following form: [DD:MM:YYYY; hh:mm:ss], and a column called "Special Event" which can only contain the strings: Positive, Negative or blank. Additionally, the file can contain as many other column headings as required by the domain. So in the case of the ICU domain this is likely to include: variables such as HR, Mean (or MAP), FiO2, SpO2, together with drugs information. Each column is typed to help TDWB spot data errors; currently only the following data types are accepted: "Timepoint", "Int" (integer), "Real" and "String". The data associated with a particular time point is held as a separate record; each record is terminated by a New Line; and files are terminated by a special terminator. The workbench has essentially 3 phases, namely: Data files (Input), Data Analysis, and Pattern Matching and Discovery which are discussed below.

"Data files" loads patient (CSV) files, performs various checks on the dataset (including: type-checking of elements, that temporal records are correctly ordered, check length of gaps between time-points), provides options for extrapolation of missing time-points etc; allows the analyst to select from all the descriptors in the CSV file which should be included in the current analysis; and set ranges for the selected descriptors. For example, the expert decided that SpO2 should have the following 5 ranges: L4 (Low-4), L3, L2, L1 & N (Normal) see Figure 1. Multiple patient datasets can

[4] Where A@T+0 refers to observation A occurs at time point 0

be loaded. There are also facilities to display the datasets in different formats: raw/original, cleaned-up (i.e., when extra and missing elements/time-points are dealt with), continuous data with a predefined set of ranges for each descriptor, and discrete where the names of the ranges are displayed. Once these processes have been successfully completed, the analyst is given the opportunity to save this information to a *project* file so that the "set up" work does not need to be done again.

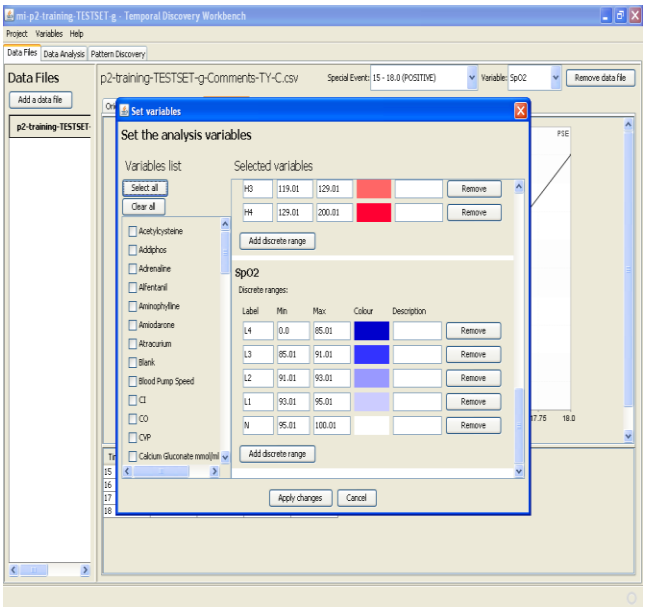

Figure 1: Showing the screen which allows the analyst to select descriptors to be used and showing how to set up the several ranges for the SpO2 Descriptor.

"Data Analysis" is not very highly developed as yet.

The *"Pattern Matching and Discovery"* module provides the most extensive set of facilities. In this summary, due to space limitations, we describe just the relevant subset of TDWB's functionality. The pattern creating modules allow the analyst to select the segments, descriptors, and descriptor ranges that are to be used in a run of the pattern generation algorithm. Additionally the analyst is able to decide whether elementary[5] or composite elements are to be the building "blocks" for the temporal patterns, the minimum and maximum number of temporal elements to be included in each pattern, and the maximum number of gaps to be included in each temporal pattern.

Another very important parameter used by the Pattern Generation algorithm is the "Positive Threshold" parameter which specifies the number of PSEs (i.e., segments that have as their last element a Positive Special Event marker) which should be matched by any pattern generated. Ideally as a result of the pattern generation process we will end up with a small number of patterns which cover all the PSEs and none of the NSEs (i.e., segments that have as their last element a Negative Special Event marker); in most real-world situations where data is noisy this is unlikely to be the case. Early on we made a design decision to make the processes of Pattern Generation and the determination of pattern "Coverage" distinct modules. There are several reasons for doing that: firstly, the processes are then much more transparent to the domain expert, and secondly if one wishes later to implement, e.g., a more sophisticated coverage algorithm one needs only add this new algorithm to the coverage module, and the Pattern Generation module is unchanged. In all studies to date we have set the positive threshold parameter to 1, so that the Pattern Generation modules report the various patterns which are found for *each* of the PSEs, and the analyst (with some support from the WB) then selects, in the Coverage module, a set of patterns which satisfy, as best as it can, the particular trade-offs, the analyst wishes to apply between covering all PSEs and no NSE.

4. TDWB: Case Study of Myocardial Damage Analysis / Prediction

In section 1, we summarized the model which we formulated as a result of several knowledge acquisition sessions with 2 domain experts, and we reported the results which that model achieved when it was applied to the test set drawn from the 51 patients (a relatively high True Positive (TP) rate (75.8%), [Sleeman et al, 2011].) What needs to be stressed here is that this model reports an association between the CVD sequences and a raised troponin value, i.e., a positive correlation is recorded if the CVD sequence occurs either *before* or *after* the raised troponin provided these events are within the defined time window of 72 hours. However, being able to predict that a CVD sequence is always/frequently followed by a raised troponin value is of course much more useful clinically. We have since run this expert model to determine how effective their

[5] **Definitions**: *Elementary Patterns* extend over a single time-period, and are of the following form: SpO2[L4]@T+0 i.e., they contain a descriptor-value pair and specify a time-point (i.e., T+0). *Composite Patterns*: are ones that involve 2 or more elementary patterns associated with a *single* time-point, e.g., SpO2[L4]@T+0, HR[H3]@T+0. *Temporal Patterns* are defined as patterns which extend over one or more time-points; further they can contain gaps i.e., time-points where none of the "active" descriptor-value pairs occur. An example of a temporal pattern is: SpO2[L4]@T+0, GAP, HR[H3]@T+2. Additionally, the elementary and composite patterns mentioned above are also valid temporal patterns.

model is at prediction [Moss et al, 2012]; prediction is the focus of the analyses we have undertaken with TDWB. Also we should point out that we are reporting the results of fewer positive and negative segments as TDWB's loading module found a number of inconsistencies in some patient datasets which had previously not been detected, and because with these studies we chose to train the system with 2/3 of the data-set (34 patients), and to use the remaining 17 patients as the test set. (This test set consists of 13 PSES and 9 NSEs.) Below we summarize the various descriptors used with each of the studies (study number is given in the first column of the table):

Study-1: This uses TDWB to run the "manual" model obtained from the clinicians over the common test dataset for these studies. Here the base patterns are: SpO2[L4], HR[L4], HR[H4], MAP[L4], or MAP[H4]; 2 of the above descriptors at level-3 (i.e., L3 or H3); 3 of the above descriptors at level-2 (i.e., L2 or H2); and FiO2[H4].[6] Also following the initial model, each reported pattern must have one of the above sub-patterns occurring at 3 out of 5 time-points (that is the model allows up to 2 gaps in each of the patterns); and this must be followed within 72 hours by a PSE (i.e., a raised troponin value).

Study-2: The remaining studies have used the (degenerate)[7] Apriori algorithm to create temporal patterns; and in all remaining studies we have just used 3 descriptors: SpO2, HR, and MAP; in all these studies the algorithm could, if supported by the data, suggest *composite* patterns. In the case of study-2 we excluded from the analysis descriptor-ranges which were N (Normal) and those at level-1 (i.e., L1 & H1). Here we specify that the minimum number of elements (elementary or composite patterns) must be 1, and the maximum number of elements must be 3. Further we specified that temporal patterns can include up to 2 gaps, so the length of the temporal patterns produced are between 1 and 5 units.

Study-3: All the parameters are the same for those in Study-2 except that descriptor-ranges at level-2 (i.e., L2 and H2) are also excluded.

Study-4: All the parameters are the same for those in Study-3 except that descriptor-ranges at level-3 (i.e., L3 and H3) are also excluded.

Study-22: All the parameters are the same as for Study-2 except that now we specify that the minimum and maximum number of elements in a temporal pattern must be 3. But up to 2 gaps are still possible, and so the temporal patterns produced here can be between 3 and 5 time-units in length. (Whereas those produced in Study-2 can between 1 and 5 units in length.)

Study-23: All the parameters are the same as for Study-3 except that now we specify that the minimum and maximum number of elements in a temporal pattern must be 3.

Study-24: All the parameters are the same as for Study-4 except that now we specify that the minimum and maximum number of elements in a temporal pattern must be 3.

Note: The patterns effectively predict that a raised troponin will be detected within 72 hours of the CVD described by the several temporal patterns produced.

Comments on Patterns produced by each of the Studies.

The first column in Table 1 gives the study number i.e., Study-1 to Study-24. The "All" column reports the number of patterns created by the Apriori algorithm for that study (with the descriptor-range pairs specified above). Because there are often a sizable number of patterns we have implemented a facility by which the Coverage module is able to select for each PSE the N highest ranked patterns. So "ALL1" corresponds to the patterns selected when the algorithm is retaining just the *top* ranked pattern for each PSE or all such patterns if a set of patterns are given equivalent ranking. The ranking of patterns is done by assigning a positive value for each PSE matched by a pattern, and a negative value corresponding to each NSE matched. So in the case of this study, the domain expert suggests a +5 and -2 respectively; note these values are parameters and can be changed for each analysis. As you can see from looking over the figures this filter is quite effective at reducing the number of patterns to be considered (in the case of Study-2 the reduction is from 318 to 15). The 4th column provides the usual metrics (True Positive (TP)/False Negative (FN)/True Negative (TN)/False Positive (FP)) for both the All and All1 sets of patterns. As mentioned earlier the role of the Coverage module is to help the analyst/domain expert select patterns; one common objective is to cover as many of the PSEs as possible, and as few NSEs as possible. (See figure 2.) This, in general, is clearly a complex optimization process, so for the moment we report the results for "All1" and for the "AllPSE+MinNSE" strategy which covers the *minimum* number of NSEs consistent with retaining the *maximum* number of PSEs. The table gives the number of resulting patterns for this strategy and the associated metrics. The final column reports, for both

[6] This is a slight simplification of FiO2's role in the model as specified by the domain experts.

[7] The degenerate Apriori algorithm is a combinatorially simpler version which is applicable only when one is seeking patterns which cover a single PSE (Positive Segment)

the All1 and AllPSE+MinNSE strategies, evaluations of the scoring function: number of TPs * 5 – number of FPs * 2; i.e., the same parameters used in the COVERAGE module to assess the "strength" of a pattern. (Recall these numbers are parameters provided at each run by the analyst.)

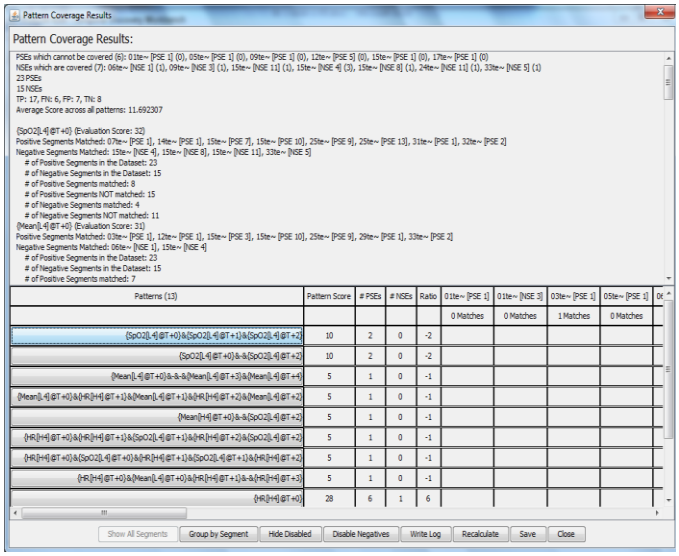

Figure 2: A Screenshot of the Coverage module.

Analysis of the Studies and Discussion

- The number of patterns reported for the Study-4 and Study-24 are both very low, this is because here TDWB used a very small number of descriptor-value pairs; in fact only those at level-4 (a total of just 5 descriptor-values pairs)
- On the other hand we see that a larger number of descriptor-value pairs results in many more patterns being produced (Study-2 and Study-22). Study-22 produces a smaller number of patterns than Study-2, as the length of the temporal patterns created here is more restricted.
- The relatively large number of patterns produced by Study-2 when, compared with Study-3, results in a higher proportion of FPs being produced.
- The previous points suggest that if the description space is too restrictive the coverage of the PSEs (i.e., TPs) is low, however if the description space is too large then many more PSEs will be covered but so will many of the NSEs. Studies 2, 3, 22, & 23 show a trade-off between these factors.
- The Strategy AllPSE+MinNSE is generally effective in reducing the number of FPs whilst retaining the maximum number of TPs.
- What is of most significance here is that the coverage produced by Studies-2, 3, 22, & 23 are all better than that produced by Study-1, even though Study-1 uses an additional descriptor, namely FiO2.[8] The differences between Study-1 and the other studies have been shown to be statistically significant.

Table 1: Summary of results for the several studies run with TDWB.

St	All #	All 1 #	Metrics (TP/FN /FP/TN) for All & All1 (same)	AllPSE+ MinNSE #	Metrics (TP/FN/ FP/TN)	Overall Score for All1 & AllPSE+ MinNSE (max 65)
1	127	6	7/6/3/6	4	7/6/2/7	29 | 31
2	318	15	13/0/9/0	4	13/0/2/7	47 | 61
3	115	13	11/2/7/2	2	11/2/3/6	41 | 49
4	16	5	8/5/5/4	4	8/5/4/5	30/32
22	312	14	12/1/9/0	4	12/1/0/9	42 | 60
23	106	12	11/2/7/2	2	11/2/3/6	41 | 49
24	7	1	3/10/1/8	1	3/10/1/8	13 | 13

Discussion: This last point reports that the Apriori (AP) algorithm, which makes systematic searches through the data, produces larger number of patterns than the model formulated as a result of knowledge capture (KC) with the domain experts. However, many of the patterns produced by the AP algorithm might not be clinically acceptable i.e., they might describe patterns which are considered by domain experts *unlikely* to precede a raised troponin value. This point needs to be investigated thoroughly.

Pattern Evaluation by Experts[9]

We asked a clinician to review each of the 13 patterns produced in Study-3, and to say whether each was **l**ikely, unlikely or not-enough-information-to-decide, and to give his reasons. We also asked him to review each of the patterns which he had classified as likely,

[8] Note the identified patterns reported here are able to predict the occurrence of PSEs, whereas in the Glasgow study (section 1), the patterns report *associations* between the identified descriptors and the PSE marker. (So in the case of association the order of the 2 entities is not significant, whereas in the case of prediction it is.)

[9] These expert-led evaluations are comparable to work in machine learning (ML) which seeks to evaluate inferences produced by ML algorithms.

and say how they could be simplified yet retained the "likely" classification. The expert indicated that 12 of the patterns were likely, and there was not enough information to decide on one pattern. We then derived the following sets of patterns: a) the remaining 12 "likely" patterns when the metrics for Study-3 did not change, and b) the 12 "likely" patterns and a further set of (12) "equivalent" patterns suggested by the expert. With this enhanced pattern set the overall metrics (TN/FN/FP/TN) for the All1 selection process improved from 11/2/7/2 to 12/1/8/1.

This preliminary study suggests that some of the types of patterns which had been excluded in the initial manual model are acceptable to this clinician. Moreover, the patterns produced as a result of the expert's selection still have a significantly broader coverage than those achieved by the "manual" model. The differences between Study-1 and the above sets of "derived" patterns have still been shown to be statistically significant.

5. Further Work

- Update the study document as a result of the pilot study, and then repeat the study with at least 3 experts. Have at least 2 analysts review the outcomes and come to a consensus on the decisions made by each of the experts. (Compare the "coverage" of the patterns produced by TDWB with the sets available after the expert evaluations.)
- Run the same type of study but with a much larger set of descriptors; including a number of treatment descriptors (eg a range of drugs) and more patient-orientated descriptors including temperature, and volume of urine output.
- Use the information about equivalence of derangements, provided in an earlier study [Sleeman et al, 2009], to generalize over the "negative" filters acquired in this study.
- Link TDWB to appropriate ontologies to provide at least the domain terminology
- Use TDWB with further clinical datasets (e.g., the onset of diabetes, when to ventilate ICU patients); as well as ones from Ecology and Finance.

Acknowledgements

- TDWB was implemented by Sam Cauvin and Michael Gibson (University of Aberdeen) with financial support from the University of Aberdeen Development Trust
- Dr Malcolm Sim, Consultant Southern General Hospital Glasgow, for carrying out the pilot study on patterns produced by TDWB.
- Useful discussions on aspects of the design of TDWB with Dr Wamberto Vasconcelos (University of Aberdeen).

This work was an extension of the routine audit process in Glasgow Royal Infirmary's ICU; requirements for further Ethical Committee Approval have been waved.

REFERENCES

[1] Agrawal R. and Srikant R. 1994. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB'94)* (Santiago de Chile, Chile, September 12-15, 1994). Morgan-Kaufman, San Francisco, 487-499.

[2] Combi, C., Keravnou-Papailiou, E., Shahar. Y. 2010. *Temporal Information Systems in Medicine*. Springer, Heidelberg.

[3] Laxman S. and Sastry P,S. 2006. A survey of temporal data mining. *SADHANA, Academy Proceedings in Engineering Sciences* 31, 2, (April 2006), 173-198.

[4] Moss, L., Sleeman, D., Sim, M., Kinsella, J. Using Cardiovascular Derangements to Predict Raised Troponin Levels. 1st International Workshop on Capturing and Refining Knowledge in the Medical Domain (KMED 2012) http://homepages.abdn.ac.uk/dcorsar/pages/kmed2012/papers.php

[5] Sleeman, D., L Moss, M Sim & J Kinsella (2011). Predicting Adverse Events: Detecting Myocardial Damage in Intensive Care Unit (ICU) Patients, Proceedings of KCAP 2011 Conference, Publ: AM press, pp73-80.

[6] Sleeman, D., Aiken, A., Moss, L., Kinsella, J., Sim, M. 2009. A system to detect inconsistencies between a domain expert's different perspectives on (classification) tasks. In *Advances in Machine Learning II, Studies in Computational Intelligence*. Springer Berlin / Heidelberg, 293-314.

[7] D Sleeman, L Moss, and J Kinsella (2015). "Temporal Discovery Workbench: a Case Study with ICU Patient Datasets". Research section of the HIS Conference, Glasgow, Sept 2014. http://ewic.bcs.org/category/18476