Automated Revision of Expert Rules for Treating Acute Abdominal Pain in Children

Sašo Džeroski[1,2], Giorgos Potamias[1], Vassilis Moustakis[1,3], Giorgos Charissis[4]

[1] FORTH-ICS, P.O.Box 1385, 711 10 Heraklion, Greece
[2] Department of Intelligent Systems, Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia (Email: Saso.Dzeroski@ijs.si)
[3] Department of Production and Management Engineering,
Technical University of Crete, 73100 Chania, Greece
[4] Director, Pediatric Clinic, University Hospital, Medical School,
University of Crete, Heraklion, Greece

Abstract. Decision making knowledge acquired directly from a medical expert is often incorrect and incomplete. Another source of knowledge about a decision making problem are examples of expert decisions in situations that have occurred in practice, stored in patient records of clinical information systems. Such examples can be used to revise the expert-provided knowledge, i.e., to discover and repair its deficiences. The revised knowledge performs better than the original one and often better than rules learned from examples alone. In addition, it inherits parts of the original expert knowledge and is thus easier to understand and accept for the expert. We present an application of the machine learning approach of theory revision to the problem of revising an expert-provided theory for treating children with acute abdominal pain.

1 Introduction

One of the most difficult problems in the development of intelligent systems is the construction of the underlying knowledge base. Normal knowledge acquisition can be divided into two phases: an initial phase in which a knowledge engineer extracts a rough set of rules from an expert and knowledge base refinement, in which the initial knowledge is refined to improve its performance. Performance improvement is also divided into two phases: the first establishes the correctness of the knowledge base and the second improves efficiency.

There exists a variety of machine learning tools that deal with improving the performance of a given rough set of rules. Explanation-based learning (De Jong and Mooney 1986) improves the efficiency of a correct domain theory. Theory refinement or theory revision (Ourston and Mooney 1994) deals with repairing a domain theory which is incorrect (incomplete and/or inconsistent). Examples are used to guide the theory revision process.

In this paper, we describe an application of a theory revision system to refine a knowledge base for treating children with acute abdominal pain. A rough set of rules was provided by the domain expert (G. Charissis). A set of patient records from the same domain was also available. The rough domain theory was then refined by the theory revision system NEITHER (Baffes and Mooney

1993), using a subset of the patient records available. The revised knowledge base performs much better than the original one and slightly better than rules learned from examples alone. In addition, it inherits parts of the original expert knowledge and is thus easier to understand and accept for the expert.

The remainder of the paper is organized as follows. The theory revision system used is briefly described in Section 2. Section 3 describes the medical domain of acute abdominal pain in children (AAPC). A general description is followed by a description of the patient records available and a decision making theory provided by the medical expert. Section 4 describes experiments designed to investigate the properties of the revised theory as the number of examples used for revision increases. This enables us to choose a subset of the available cases of an appropriate size, with which to perform revision. Section 5 describes the resulting revised theory and compares it to the original theory and a theory learned only from the examples used for revision. Section 6 first summarizes the contributions of the paper, then discusses related and further work.

2 The theory revision system NEITHER

The problem of theory revision (or knowledge base refinement) can be defined as follows: **Given** an imperfect domain theory (knowledge base) in the form of classification rules and a set of classified examples, **find** an approximately minimal syntactic revision of the domain theory that correctly classifies all of the examples.

A representative recent system that addresses this problem is EITHER (Ourston and Mooney 1994). EITHER refines propositional Horn-clause theories using a suite of abductive, deductive and inductive techniques. Deduction is used to identify the problems with the domain theory, while abduction and induction are used to correct them.

Two kinds of problems are encountered within imperfect domain theories: over-generality occurs when an example is classified into a class other than the correct one, while over-specificity occurs when an example cannot be proven to belong to the correct class. Note that a single example can be misclassified both ways at the same time. Overly general rules are either specialized by adding new conditions to their antecedents or are deleted from the knowledge base. Problems of over-specificity are solved by generalizing the antecedents of existing rules, e.g., by removing conditions from them, or by the induction of new rules.

The basic algorithm used by EITHER has three steps.

1. It first computes all possible repairs for each misclassified example.
2. It then enters a loop to select a subset of these repairs that can be applied to the theory to fix the misclassified examples. Repairs are ranked according to a benefit-to-cost-ratio between the number of examples fixed and the size of the repair and the number of new misclassifications it creates. The best repair is added at each iteration.
3. Finally, the selected subset of repairs is applied to the theory.

Any remaining misclassifications are solved by applying induction.

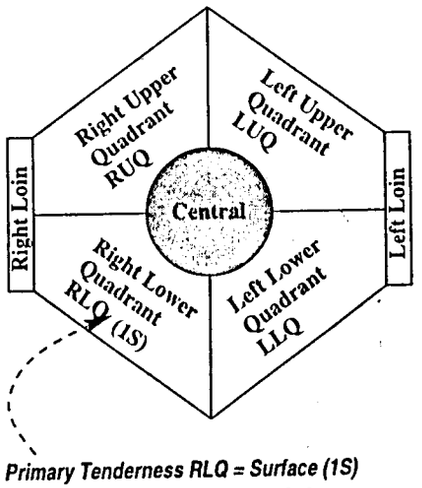

Fig. 1. A geometrical representation of the abdomen areas (de Dombal 1991).

The process of computing all possible repairs for each example is very costly: it can be exponential in the size of the theory. A new version of EITHER, called NEITHER (Baffes and Mooney 1993) has been thus implemented which adopts a greedy approach. Its main loop computes a single repair for each example, applies the best repair to the theory removing all examples fixed by the repair, and repeats until all examples are fixed. In contrast to EITHER, NEITHER's algorithm is linear in the size of the theory.

In our experiments, we used the theory revision system NEITHER. Both EITHER and NEITHER can introduce intermediate concepts during the revision process. NEITHER, however, has an option to avoid the introduction of new intermediate concepts, which we employed in our experiments.

3 Acute abdominal pain in children

The domain of acute abdominal pain in children (AAPC) encompasses a set of symptoms that cause severe pain, discomfort and increased tenderness in the abdomen of the child. AAPC originates from disorders either in the intra-abdominal or the extra-abdominal areas (Waldschmit and Charissis 1990).

Management of patients is based on an explicit protocol by de Dombal (1991) that captures pain specifics, related symptoms and results of laboratory tests. The attending physician needs to diagnose the cause of pain and then make one of the following mutually exclusive treatment decisions: 1) discharge the child (in case the cause of the pain is not pathologic), 2) proceed to immediate operation, or 3) follow-up the case for a period of six to eight hours at the end of which patient condition is re-assessed and the child is either discharged or admitted for operation.

In case an operation decision is adopted, the physician should already have in mind a spectrum of different potential causes which are to be confirmed or rejected and treated accordingly during the surgery operation.

Table 1. The original attributes. The short and long names of each attribute are listed.

AGE_GROUP (age)	ANALGET (on_analgetics)
P_DUR (pain_duration)	ANTIAC (on_antiacid)
P_START (pain_started)	ANTIEPIL (on_antiepileptics)
TYPE (type_of_pain)	ANTIBRO (on_antibrochial)
SEVERITY (severity_of_pain)	CHEMOTH (on_chemotherapy)
MOVE (movement)	MOOD (mood)
COUGH (coughing)	COLOR (color)
RESPIR (respiration)	PULSEQ (pulse)
EATING (eating)	D_PRESQ (diastolic_pressure)
R_R_SHOULD (rebound_right_shoulder)	S_PRESQ (systolic_pressure)
R_L_SHOULD (rebound_left_shoulder)	TEMPQ (temperature)
R_LOIN (rebound_loin)	DISTENS (distension)
SITE_{LUQ,RUQ,LLQ,RLQ,CENTRAL,LLO,RLO} (site_of_pain_at_onset_at_ {LUQ,RUQ,LLQ,RLQ,Central,LLo,RLo})	TENDER_ {LUQ,RUQ,LLQ,RLQ,CENTRAL,LLO,RLO} (primary_tenderness_ {LUQ,RUQ,LLQ,RLQ,Central,LLo,RLo})
ANOREXIA (anorexia)	REBOUND (rebound)
NAUSEA (nausea)	RIGIDITY (rigidity)
VOMITING (vomiting)	GUARDING (guarding)
BOWELS (bowels)	B_SOUNDS (bowel_sounds)
BLOOD_STOOLS (blood_in_stools)	RECTAL (rectal)
MUCUS_STOOLS (mucus_in_stools)	HT (hematocrit)
MICRUR (micrurition)	W_CELLS (white_cells)
HAEMATUR (haematuria)	NEUTRO (neutrophiles)
RESPIRATORY (respiratory)	LEUCO (leucocytes)
HEART (heart_disease)	ERYTHRO (erythrocytes)
ABDO (abdominal)	BACTERIA (bacteria)
HAEMATOL (haematologic)	PROTEIN (protein)
ANTIBIOT (on_antibiotics)	

3.1 The patient records

A total of 312 AAPC patient records have been selected from a database installed and running at the Pediatric Surgery Clinic of the University Hospital at Heraklion, Greece. These are described with 63 attributes, which are listed in Table 1. Two groups of attributes refer to the site of pain at onset and primary tenderness. Both derive from a geometrical model of the abdominal area (see Figure 1), originally introduced by de Dombal (1991), that facilitates accurate capturing of expert assessment with respect to the area of tenderness and the site of pain at onset.

Five attributes are continuous (age, pulse, diastolic pressure, systolic pressure, and temperature). These have been discretized into two intervals each according to the suggestions of the domain expert (G. Charissis). Of the 312 patients, 144 have been discharged, 80 operated, and 88 assigned for follow up after their first examination.

3.2 The expert domain theory

The medical expert has formulated a theory that encompasses his theoretical decision making knowledge about this domain. The theory, given in Table 3,

Table 2. Attributes introduced for the expert rules and their definitions in terms of the original attributes.

```
COUGH_CHANGE: YES IF (COUGHING GETTING_WORSE) OR (COUGHING GETTING_BETTER)
FEVER_OR_VOMITING: YES IF (TEMPQ GREQ_37) OR (VOMITING YES)
NAUSEA_VOMITING: YES IF (NAUSEA YES) OR (VOMITING YES)
RIGIDITY_OR_GUARDING: YES IF (RIGIDITY YES) OR (GUARDING YES)
URINE_TEST_ABNORMAL: YES IF (NOT (LEUCO NORMAL)) OR (NOT (ERYTHRO NORMAL))
                     OR (NOT (BACTERIA NORMAL)) OR (NOT (PROTEIN NORMAL))
TENDER_YES: YES IF for ANY X of [LUQ,RUQ,LLQ,RLQ,CENTRAL,LLO,RLO]
            (NOT (TENDER_X NONE))
TENDER_SHALLOW_DEEP: YES IF for ANY X of [LUQ,RUQ,LLQ,RLQ,CENTRAL,LLO,RLO]
                     (TENDER_X SHALLOW) OR (TENDER_X DEEP)
TENDER_SHALLOW: YES IF for ANY X of [LUQ,RUQ,LLQ,RLQ,CENTRAL,LLO,RLO]
                (TENDER_X SHALLOW)
TENDER_EVERYWHERE_SHALLOW_OR_DEEP:
                YES IF for ALL X of [LUQ,RUQ,LLQ,RLQ,CENTRAL,LLO,RLO]
                (TENDER_X SHALLOW) OR (TENDER_X DEEP)
TENDER_CENTRAL_SHALLOW_OR_DEEP: YES IF (TENDER_CENTRAL SHALLOW)
                                OR (TENDER_CENTRAL DEEP)
TENDER_RLQ_SURFACE_SHALLOW: YES IF (TENDER_RLQ SURFACE)
                            OR (TENDER_RLQ SHALLOW)
TENDER_LLQ_SURFACE_OR_SHALLOW: YES IF (TENDER_LLQ SURFACE)
                               OR (TENDER_LLQ SHALLOW)
TENDER_LUQ_SURFACE_OR_SHALLOW: YES IF (TENDER_LUQ SURFACE)
                               OR (TENDER_LUQ SHALLOW)
```

comprises 22 rules that prescribe patient treatment scenarios based on the 63 attributes listed in Table 1. The expert provided rules make use of 13 additional attributes, defined in terms of original attributes. These new attributes and their definitions are listed in Table 2. They have been defined in order to keep the expert provided rules in the form of conjunctions.

Each of the rules has one of the three possible treatments in the conclusion part: 6 rules recommend `operate`, 5 `discharge`, and 11 `follow up`. The antecedent of each rule is a list or conjunction of attribute value pairs. For example, rule `01` of Table 3 recommends the treatment `operate` if the attribute `RIGIDITY` has the value `YES`.

The theoretical expert knowledge encompassed in the rules of Table 3 is not entirely consistent with the treatment decisions taken in practice. In fact, when tested on the 312 cases provided, the rules prescribe a unique treatment that agrees with the actual one in only 50 cases: a correct treatment is only prescribed in 16% of the cases.

The poor performance of the rules should not be perceived as a consequence of poor expert performance. The main reason for the poor performance of the rules is that the expert is not accustomed to expressing his medical knowledge in terms of rules. For instance, with rule `01` of Table 3 the expert actually meant that the occurrence of `RIGIDITY` is a necessary, but not a sufficient condition to `operate`. This has serious implications for the effectiveness of using a rule

Table 3. The original theory as provided by the expert. Annotations indicate what happened to each rule during theory revision. The `Ri`'s refer to Table 4.

```
01: OPERATE <- (RIGIDITY YES) :specialized into R6 and R7
02: OPERATE <- (ANTIBIOT YES) (ABDO YES) (TEMPQ GR_37)
               (TENDER_RLQ_SURFACE_SHALLOW YES) :unchanged
03: OPERATE <- (TYPE STEADY) (TEMPQ GR_37) (TENDER_RLQ SHALLOW)
               (NOT (ERYTHRO NORMAL)) :unchanged
04: OPERATE <- (ABDO YES) (NAUSEA_VOMITING YES)
               (TEMPQ GR_37) (REBOUND YES) :unchanged
05: OPERATE <- (AGE_GROUP LESS_4) (TYPE STEADY) (NAUSEA_VOMITING YES)
               (P_DUR TODAY) (TEMPQ GR_37)
               (TENDER_RLQ_SURFACE_SHALLOW YES) :unchanged
06: OPERATE <- (AGE_GROUP GREQ_4) (TYPE STEADY) (NAUSEA_VOMITING YES)
               (TEMPQ GR_37) (W_CELLS INCREASED) (NEUTRO INCREASED)
               (TENDER_RLQ_SURFACE_SHALLOW YES) :unchanged
07: FOLLOW_UP <- (TEMPQ GR_37) (COUGH_CHANGE YES) :unchanged
08: FOLLOW_UP <- (NAUSEA YES) (ANOREXIA YES) (AGE_GROUP LESS_4) :deleted
09: FOLLOW_UP <- (B_SOUNDS INCREASED)
                 (TENDER_SHALLOW_DEEP YES) :specialized into R14
010: FOLLOW_UP <- (VOMITING YES) (BOWELS DIARRHEA)
                  (TENDER_SHALLOW_DEEP YES) :unchanged
011: FOLLOW_UP <- (HAEMATUR YES) (NAUSEA_VOMITING YES) :unchanged
012: FOLLOW_UP <- (W_CELLS INCREASED) (NEUTRO INCREASED)
                  (TENDER_YES YES) :specialized into R15
013: FOLLOW_UP <- (TENDER_YES YES) :deleted
014: FOLLOW_UP <- (BOWELS CONSTIPATION) :deleted
015: FOLLOW_UP <- (NAUSEA_VOMITING YES) (URINE_TEST_ABNORMAL YES) :deleted
016: FOLLOW_UP <- (NAUSEA_VOMITING YES) (MICRUR DYSURIA)
                  (TENDER_SHALLOW_DEEP YES) :unchanged
017: FOLLOW_UP <- (NAUSEA_VOMITING YES) (TENDER_SHALLOW YES) :deleted
018: DISCHARGE <- (TENDER_LUQ_SURFACE_OR_SHALLOW YES)
                  (RIGIDITY_OR_GUARDING YES) :deleted
019: DISCHARGE <- (TENDER_LLQ_SURFACE_OR_SHALLOW YES)
                  (RIGIDITY YES) (GUARDING YES) :deleted
020: DISCHARGE <- (TYPE COLICKY) (BOWELS CONSTIPATION)
                  (VOMITING YES) (TEMPQ LEQ_37) :unchanged
021: DISCHARGE <- (TENDER_CENTRAL_SHALLOW_OR_DEEP YES)
                  (FEVER_OR_VOMITING YES) :deleted
022: DISCHARGE <- (TENDER_EVERYWHERE_SHALLOW_OR_DEEP YES) (RIGIDITY NO)
                  (GUARDING NO) (VOMITING NO) (TEMPQ LEQ_37) :unchanged
```

based representation and poses challenges for further research on the interaction between a human expert and a revision system. On the technical level, there are several reasons for this poor performance. To start with, the set of rules as a whole sometimes prescribes two or even three conflicting treatments at the same time. For example, whenever rule `019` recommends the treatment `discharge`, the rule `01` will recommend the treatment `operate`. For some patients, only one inappropriate treatment is recommended. Finally, no treatment is recommended for some patients.

In any case, the expert provided rules obviously need to be revised if they are to be used for any practical purpose. In the remainder of the paper, we describe the application of the theory revision system NEITHER, described in Section 2, to revise the expert provided rules using (a subset of) the provided patient records.

4 The AAPC learning curve for theory revision

This section describes the experiments designed to investigate how the accuracy of the revised theory on unseen cases and the size of the revised theory behave as the number of examples used for revision increases. This in order to enable us to choose an appropriately sized subset of the available patient records for revision.

The first experiment was designed as follows. Ten different partitions of the 312 cases into a training R_i and testing Ei ($i = 0..9$) set (sized roughly 281 and 31 cases, respectively) were created using the ten-fold cross-validation method. Stratified cross-validation was used, making an effort to preserve the relative proportions of treatment decisions from the entire set of cases. For example, a training set R_i would have roughly 72 ($281 * 80/312$) cases to which treatment `operate` was prescribed.

For each of the ten partitions, the following actions were taken: ten different subsets of R_i were created such that $R_{i0} \subset R_{i1} \subset \ldots \subset R_{i9} = R_i$, where R_{ij} is of size approximately $(j + 1) * 0.1 * |R_i|$. Again, an effort was made to preserve the relative proportions of treatment decisions. The original theory from Table 3 was revised using NEITHER and the cases from R_{ij} yielding a revised theory T_{ij}. T_{ij} was then tested on the unseen cases from Ei and its accuracy A_{ij} recorded. The accuracies A_{ij} are averaged over the ten splits to yield A_j. It is these averages that are shown in Figure 2a): the first point represents the accuracy of the original theory, while the last point represents A_9, the accuracy of the theory revised with approximately 280 examples.

In the second experiment, the behavior of the size of the revised theory was investigated. Ten different partitions of the entire dataset R were created into training sets P_i and testing sets Q_i ($i = 0..9$), such that $P_i \cup Q_i = R$, $P_0 \subset P_1 \subset \ldots \subset P_9 = R$ and $|P_i| \approx (i+1) * 0.1 * |R|$. Again, an effort was made to preserve the relative proportions of the three different treatment decisions as in the entire dataset.

The original theory from Table 3 was revised using NEITHER and the cases from P_i to yield a revised theory S_i. The sizes of these theories, measured as $|S_i|$, i.e., the number of rules in S_i, are shown in Figure 2b).

To summarize the results of the two experiments, as the number of examples used for revision increases, the accuracy on unseen cases increases at first, but levels off quickly at roughly 100 examples. The size of the revised theory, however, grows more or less monotonically with the number of examples used for revision. The latter result suggests that we should use as small a set of cases for revision as possible, provided this is not at the expense of the accuracy of the revised theory. As accuracy levels off at 100 examples, this seems to be an appropriate number of cases to use for theory revision in the AAPC domain.

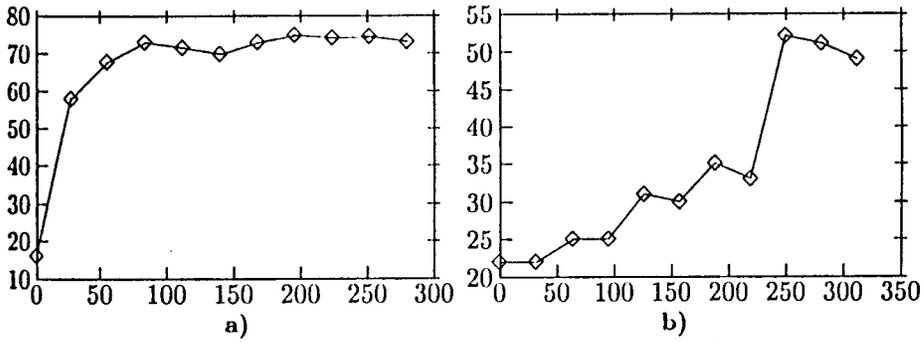

Fig. 2. Learning curves for theory revision in the domain of acute abdominal pain in children. The x-axes show the number of examples used to revise the theory provided by the expert. For graph a) the y-axis shows the accuracy of the revised theory on unseen cases. For graph b), the y-axis shows the number of rules in the revised theory.

5 The revised AAPC theory

Having determined the appropriate number of examples to perform revision with, we conducted the following experiment. The whole dataset of 312 samples was split into a training/revision set of size 30% (95 examples) and a testing set of 217 examples. The relative proportions of the treatment decisions were preserved: the training set contains 24 `operate` cases, 26 `follow up` and 45 `discharge` cases. The original theory (Table 3) was then given to NEITHER to revise using the 95 cases of the revision set.

The revised theory produced by NEITHER is given in Table 4. NEITHER was also given the empty theory to revise, which causes it to construct a theory in an entirely inductive way, i.e., by learning from the 95 examples only. This learned theory is shown in Table 5.

The accuracies of all three theories (original, revised and learned from examples only) were measured on both the training/revision set and on the unseen cases from the testing set and are shown in Table 6. The original theory performs poorly on both the training and the testing set (19% and 15%, respectively). After revision, its performance is drastically improved. It is now 100% accurate on the training set and 75% accurate on the unseen cases. The theory learned from examples only is 100% accurate on the training cases and 73% on the testing cases. Thus, the revised theory performs much better than the original theory and slightly better than the theory learned from examples only.

To highlight the effects of theory revision on the original theory, the rules in the original theory and the revised theory have been labeled and annotated. For example, rule `01` of the original theory was specialized, resulting into rules `R6` and `R7` of the revised theory. Rule `02` has been left unchanged as rule `R1` of the revised theory, and rule `08` has been deleted. On the other side, rule `R11` of the revised theory is a copy of rule `016`, rules `R12` and `R13` are entirely new, while rule `R14` is a specialization of the rule `09`.

Table 4. The revised theory. The comments provided indicate the origin of each rule.

```
R1: OPERATE <- (ANTIBIOT YES) (ABDO YES) (TEMPQ GR_37)
               (TENDER_RLQ_SURFACE_SHALLOW YES) :unchanged 02
R2: OPERATE <- (TYPE STEADY) (TEMPQ GR_37) (TENDER_RLQ SHALLOW)
               (NOT (ERYTHRO NORMAL)) :unchanged 03
R3: OPERATE <- (ABDO YES) (NAUSEA_VOMITING YES)
               (TEMPQ GR_37) (REBOUND YES) :unchanged 04
R4: OPERATE <- (AGE_GROUP LESS_4) (TYPE STEADY) (NAUSEA_VOMITING YES)
               (P_DUR TODAY) (TEMPQ GR_37)
               (TENDER_RLQ_SURFACE_SHALLOW YES) :unchanged 05
R5: OPERATE <- (AGE_GROUP GREQ_4) (TYPE STEADY) (NAUSEA_VOMITING YES)
               (TEMPQ GR_37) (W_CELLS INCREASED) (NEUTRO INCREASED)
               (TENDER_RLQ_SURFACE_SHALLOW YES) :unchanged 06
R6: OPERATE <- (RIGIDITY YES) (TYPE STEADY) (SITE_RUQ NO) :specializes 01
R7: OPERATE <- (RIGIDITY YES) (TENDER_RLQ SURFACE) :specializes 01
R8: FOLLOW_UP <- (TEMPQ GR_37) (COUGH_CHANGE YES) :unchanged 07
R9: FOLLOW_UP <- (VOMITING YES) (BOWELS DIARRHEA)
                 (TENDER_SHALLOW_DEEP YES) :unchanged 010
R10: FOLLOW_UP <- (HAEMATUR YES) (NAUSEA_VOMITING YES) :unchanged 011
R11: FOLLOW_UP <- (NAUSEA_VOMITING YES) (MICRUR DYSURIA)
                  (TENDER_SHALLOW_DEEP YES) :unchanged 016
R12: FOLLOW_UP <- (TENDER_YES NO) :new
R13: FOLLOW_UP <- (TEMPQ GREQ_37) (COLOR NORMAL)
                  (TENDER_SHALLOW_DEEP YES) (COUGH GETTING_WORSE) :new
R14: FOLLOW_UP <- (B_SOUNDS INCREASED)
                  (TENDER_SHALLOW_DEEP YES) (SITE_LUQ NO) :specializes 09
R15: FOLLOW_UP <- (W_CELLS INCREASED) (NEUTRO INCREASED) (TENDER_YES YES)
                  (SITE_CENTRAL NO) (RIGIDITY NO) :specializes 012
R16: FOLLOW_UP <- (TENDER_SHALLOW YES) (TENDER_LLQ NONE)
                  (SITE_RUQ YES) :new
R17: FOLLOW_UP <- (TENDER_SHALLOW YES) (TENDER_LLQ NONE)
                  (P_DUR DAYS_1_3) :new
R18: FOLLOW_UP <- (TENDER_SHALLOW YES) (TENDER_LLQ NONE) (PULSEQ LESS_100)
                  (RIGIDITY NO) :new
R19: DISCHARGE <- (TYPE COLICKY) (BOWELS CONSTIPATION)
                  (VOMITING YES) (TEMPQ LEQ_37) :unchanged 020
R20: DISCHARGE <- (TENDER_EVERYWHERE_SHALLOW_OR_DEEP YES) (RIGIDITY NO)
                  (GUARDING NO)(VOMITING NO)(TEMPQ LEQ_37) :unchanged 022
R21: DISCHARGE <- (TENDER_RUQ DEEP) (P_START SUDDENLY) :new
R22: DISCHARGE <- (TENDER_CENTRAL_SHALLOW_OR_DEEP YES) (SITE_LUQ YES) :new
R23: DISCHARGE <- (TENDER_RLQ_SURFACE_SHALLOW NO) (SITE_RUQ NO)
                  (PULSEQ GREQ_100) (SITE_LLQ YES) :new
R24: DISCHARGE <- (TENDER_RLQ_SURFACE_SHALLOW NO) (SITE_RUQ NO)
                  (NEUTRO NORMAL) (TENDER_LUQ NONE)
                  (RESPIR GETTING_WORSE) :new
R25: DISCHARGE <- (TENDER_RLQ_SURFACE_SHALLOW NO) (SITE_RUQ NO)
                  (NEUTRO NORMAL) (TENDER_LUQ NONE) (COUGH NO_CHANGE) :new
```


Table 7 gives several statistics on the numbers of rules in the three theories and also summarizes the changes made to the original theory during the revision process. The first four rows of the table pertain to the original theory, the next four to the revised and the last row to the theory learned from examples only. The latter is the shortest with 15 rules, the revised the longest with 25 rules.

Table 5. The theory learned from examples only.

```
L1: OPERATE <- (RIGIDITY YES) (TYPE STEADY) (SITE_RUQ NO)
L2: OPERATE <- (TENDER_RLQ SURFACE)
L3: FOLLOW_UP <- (TENDER_SHALLOW YES) (NEUTRO INCREASED) (TYPE COLICKY)
L4: FOLLOW_UP <- (W_CELLS INCREASED) (TENDER_SHALLOW_DEEP YES)
                 (TENDER_SHALLOW NO) (ANTIBRO NO)
L5: FOLLOW_UP <- (PULSEQ LESS_100) (D_PRESQ GREQ_120) (SITE_LUQ YES)
L6: FOLLOW_UP <- (PULSEQ LESS_100) (W_CELLS INCREASED) (NAUSEA NO)
L7: FOLLOW_UP <- (SITE_RUQ YES) (TENDER_LLQ NONE) (TYPE COLICKY)
L8: FOLLOW_UP <- (TENDER_RLQ SHALLOW) (ERYTHRO RARE)
L9: FOLLOW_UP <- (TENDER_SHALLOW YES) (TENDER_LLQ NONE)
                 (TYPE INTERMITTENT)
L10: DISCHARGE <- (TENDER_CENTRAL DEEP) (COUGH NO_CHANGE)
L11: DISCHARGE <- (ANTIBRO YES)
L12: DISCHARGE <- (TENDER_RLO SURFACE)
L13: DISCHARGE <- (TENDER_RLQ_SURFACE_SHALLOW NO) (TENDER_SHALLOW NO)
                  (W_CELLS NORMAL) (TENDER_RLQ DEEP)
L14: DISCHARGE <- (NEUTRO NORMAL) (TENDER_LLQ SHALLOW)
L15: DISCHARGE <- (W_CELLS NORMAL) (SITE_RUQ NO) (TEMPQ LESS_37)
```

We can see that the revised theory has three rules more than the original theory. In the original theory, half of the rules were left unchanged during the revision process, 8 rules were deleted (of which none recommended `operate`, 5 recommended `follow up`, and 3 recommended `discharge`), and 3 were specialized (resulting in 4 rules in the revised theory). No new rule was created for recommendation `operate` in the revised theory, while 5 new rules were created for each of `follow up` and `discharge`. The specialist provided rules recommending `operate` are thus much more consistent with actual clinical practice than is the case with the other two recommendations.

Table 6. The accuracies of the original, revised and learned theory on the training and testing sets of examples.

Accuracy of _ theory on _	Training set (95 examples)	Testing set (217 examples)
Original	19%	15%
Revised	100%	75%
Learned	100%	73%

Finally, let us consider the number of attributes that appear in the three theories. The revised theory uses 37 attributes, of which 25 are used in the original theory (which uses 33 attributes): 8 attributes are thus deleted from the original theory and 12 other added during the theory revision process. The theory learned from examples only uses 20 attributes, of which 10 are used in the original theory and 14 in the revised theory.

Table 7. Statistics on the number of rules of the original, revised and learned theory.

Number of rules	Total	Operate	Follow up	Discharge
Original	22	6	11	5
unchanged	11	5	4	2
deleted	8	0	5	3
specialized	3	1	2	0
Revised	25	7	11	7
old	11	5	4	2
specializations	4	2	2	0
new	10	0	5	5
Learned	15	2	7	6

Nine attributes are used in all three theories. One might say that these 9 attributes are essential for appropriate treatment of children with acute abdominal pain. The 9 attributes are as follows: type of pain, temperature, primary tenderness at right left quadrant (RLQ), rigidity, white cells, neutrophiles, erythrocytes, tenderness shallow or deep (anywhere), and tenderness surface or shallow at RLQ. This is in agreement with existing expert knowledge.

6 Discussion

We have presented an application of theory revision techniques from the field of machine learning in the domain of acute abdominal pain in children (AAPC). The expert provided domain knowledge turns out to disagree with actual cases that appear in clinical practice. We have first determined the appropriate number of cases to be used for revision and then revised the expert provided domain knowledge using patient records. The revised knowledge performs much better than the original knowledge and slightly better than the knowledge learned from patient records alone. In addition, the revised knowledge inherits parts of the original expert knowledge and is thus easier to understand and be used in practice by the expert.

One of the reasons for the ppor performance of the initial set of rules provided by the expert is the fact that they have not been validated prior to evaluating their performance. If they had been validated, the improvement in performance may have been less drastic. However, this rises the point that theory revision can be used to both validate and refine a set of rules provided by an expert.

Potamias et al. (1996) have preliminary applied theory revision techniques in the AAPC domain. The expert provided theory there consists of the `operate` and `follow up` rules of the theory considered here. The interactive knowledge revision tools of MOBAL (Morik et al. 1993) are used to revise the restricted domain theory. Revisions to the theory have to be approved by a human supervisor of the revision process, who has also to provide rule schemata (templates) that the revised rules have to match. MOBAL allows a first-order logic representation of rules and second-order logic representation of rule templates.

There are obviously arguments for and against each of the two approaches. NEITHER uses a propositional representation and requires essentially no human intervention during the revision process. It is consequently faster and easier to apply. MOBAL, on the other hand, uses a more powerful representation and allows tight control of the revision process by a human supervisor. It is consequently much slower and more difficult to apply. In any case, the experiments conducted in this study are more complete, as a theory with rules for all three decisions is used and revised. In further work, we plan to apply MOBAL and other theory revision systems that use more powerful formalisms, such as FORTE (Richards and Mooney 1991) on the problem studied here and compare the advantages and disadvantages of each approach.

Acknowledgements

At the time of writing this paper, S. Džeroski was supported through an ERCIM fellowship by the Foundation of Research and Technology - Hellas. This work was partially supported by the research grant PENED 1248 from the Ministry of Development of Greece (V. Moustakis). The AAPC database development and maintenance was done by the Pediatric Clinic, University Hospital, Heraklion, Greece (G. Charissis), in the context of the STAR 1045 European Health Telematics project. Thanks go to an anonymous referee for the useful comments provided.

References

1. Baffes, P.T., and Mooney, R.J. (1993). Symbolic revision of theories with M-of-N rules. In *Proc. Thirteenth International Joint Conference on Artificial Intelligence*, pp. 1135-1140. Chambery, France, 1993.
2. de Dombal, F.T. (1991). *Diagnosis of abdominal pain.* Churchil Livingstone, Singapore.
3. De Jong, J.F., and Mooney, R.J. (1986). Explanation based learning: an alternative view. *Machine Learning*, 1(2): 145–176.
4. Morik, K., Wrobel, S., Kietz, J.-U., and Emde, W. (1993). *Knowledge Acquisition and Machine Learning - Theory, Methods, and Applications.* Academic Press, London.
5. Potamias, G., Moustakis, V., and Charissis, G. (1996). Iterative knowledge construction and maintenance. In *Proc. ICML'96 Workshop Machine Learning Meets Human Computer Interaction.* Bari, Italy, 1996.
6. Richards, B.L., and Mooney, R.J. (1991). Refinement of first-order Horn-clause domain theories. *Machine Learning*, 19(2): 95–131.
7. Ourston, D., and Mooney, R.J. (1994). Theory refinement combining analytical and empirical methods. *Artificial Intelligence*, 66: 273–309.
8. Waldschmit, J., and Charissis, G. (1990). *Das akute Abdomen im Kindesalter. Diagnose und Differentialdiagnose.* Edition Medezin, New York.