Description Logics with Approximate Definitions
Precise Modeling of Vague Concepts

Stefan Schlobach and **Michel Klein**
Department of Artificial Intelligence
Vrije Universteit Amsterdam
{schlobac,michel.klein}@few.vu.nl

Linda Peelen
Department of Medical Informatics
Academic Medical Center, Amsterdam
l.m.peelen@amc.uva.nl

Abstract

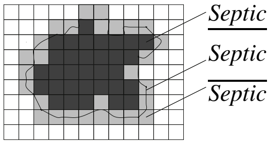
We extend traditional Description Logics (DL) with a simple mechanism to handle approximate concept definitions in a qualitative way. Often, for example in medical applications, concepts are not definable in a crisp way but can fairly exhaustively be constrained through a particular sub- and a particular super-concept. We introduce such **lower** and **upper approximations** based on rough-set semantics, and show that reasoning in these languages can be reduced to standard DL satisfiability. This allows us to apply *Rough Description Logics* in a study of medical trials about sepsis patients, which is a typical application for precise modeling of vague knowledge. The study shows that Rough DL-based reasoning can be done in a realistic use case and that modeling vague knowledge helps to answer important questions in the design of clinical trials.

1 Introduction

Many existing knowledge modeling techniques are best suited for modeling crisp knowledge. In practice, however, it is not always possible to make clear-cut distinctions. A modeler frequently has to account for borderline cases. Approaches that do take such uncertainty or vagueness into account often do this via some kind of weighting mechanism or an approach based on fuzzy sets. A drawback of these approaches is that uncertainty is introduced in the model, which often has the consequence that no crisp answers can be given to queries on the model. This paper introduces a complementary mechanism that allows for modelling of vague knowledge by crisp specification of approximations of a concept.

Medicine is a typical domain where concepts cannot always be described in a crisp manner. E.g., the definition of a disease is not always clear-cut, especially if a single marker is lacking that distinguishes a patient with a disease from a patient without the disease. This is common in psychiatry and in diseases in which the underlying pathology of the disease is unclear. An example of the latter is *sepsis*. Rough Description Logics (Rough DL) provides us with the possibility to describe such diseases for which a crisp definition is lacking.

Rough DL extends classical Description Logic ([Baader *et al.*, 2003]) by two modal-like operators, called the lower and upper approximations. In the spirit of Rough Set theory [Pawlak, 1982], two concepts approximate an underspecified, vague, concept as particular sub- and super-concepts, describing *which elements* are **definitely**, respectively **possibly**, elements of the concept. The following picture illustrates the general idea:



Each square denotes a set of domain elements, which cannot further be discerned by any available criterion. Then, the circled line denotes the set of septic patients, i.e., the vague concept which we are incapable to formally define. If we capture this lack of criteria to discern between two objects as a indiscernibility relation dis$^{\sim}$, we can formally define the upper approximation as the set of patients that are indiscernible from at least one septic patient.

$$\overline{\text{Septic}} \equiv \{pat_1 \mid \exists\, pat_2\text{: dis}^{\sim}(pat_1,pat_2)\ \&\ pat_2 \in \text{Septic}\}.$$

Similarly, we can define the lower approximation as the set of patients containing all, and only those patients, for which it is known that all indiscernible patients must be septic.

$$\underline{\text{Septic}} \equiv \{pat_1 \mid \forall\, pat_2\text{: dis}^{\sim}(pat_1,pat_2) \rightarrow pat_2 \in \text{Septic}\}$$

In our picture, the upper approximation is depicted as the union of the dark squares (the lower approximation), and the gray squares, the boundary. This semantics can be transferred to Rough DL approximations in a straightforward way: the patients in the concept $\underline{\mathit{Septic}}$ are the **definitely** septic patients, those that are unmistakably septic, the concept $\overline{\mathit{Septic}}$ models the **possibly** septic patients, as opposed to the white squares, which model **definitely not** septic patients. These approximations are to be defined in a crisp way.

Technically, Rough DL are very simple languages, as they can be simulated with traditional DL without added expressiveness. This means that reasoning can be performed by translation, and subsequent use of a common DL reasoner. We consider it a big advantage of our approach that we can use an optimised DL reasoner without having to develop new ad-hoc decision procedures and implementations. In other words, our Rough DL's are strictly speaking not more expressive than traditional DL's, but the notions that we introduce are useful modeling devices for specific types of knowledge (namely non-crisp concepts).

Our current research was motivated by a recent study of the definitions for sepsis used in clinical trials. Before a medical treatment can be used in daily clinical practice, its effect and impact on the patient have to be investigated in a clinical trial. When several trials have been performed it is interesting to compare the results of those trials. Unfortunately, the nine different trials that were investigated in [Peelen *et al.*, 2005] showed too much variation in their definitions of severe sepsis patients to enable a fair comparison of trial results.

We show how to use Rough DL to formalise and compare sepsis definitions used in different trials. Describing sepsis through approximations enforces powerful semantic consequences. Rough DL turns out to be an appropriate logical representation language to model vague concepts and provide crisp answers to queries, and can thereby assist in the *validation* of existing and, ultimately, the *construction* of new trials.

The remainder of the paper is structured as follows. First, we introduce our use-case, the medical condition sepsis. In Section 3, Rough DL is defined as an extension to standard DL for modeling vague knowledge. We give some logical consequences of the semantics of the extension, and explain how reasoning can be done by reducing Rough DL to standard DL reasoning. In Section 4, we use Rough DL to model definitions of severe sepsis used in different clinical trials. Based on real patient data we evaluate the design of the trials.

2 Sepsis: a condition with a vague definition

Severe sepsis is our example for vague information throughout the paper. Therefore, we will briefly provide some medical background. Sepsis is a disease in which the immune system of the patient overreacts to an infection. Due to this reaction the patient becomes severely ill, which easily results in organ failure and eventually death. The cause and underlying cellular pathways of this disease are unclear, which hinders the precise characterization of the sepsis patient. Therefore, a *consensus definition* of sepsis was established in 1992 to define several stages of sepsis [Bone, R.C., 1992]. This definition does not provide a precise definition of sepsis, but gives the criteria for which there was a consensus that they should at least hold for a patient with severe sepsis. In this paper we focus on the patients with *severe sepsis*, but for brevity we will refer to these patients as *septic*. The consensus statement defines patients with severe sepsis as 'patients having a confirmed infection with at least two out of four Systemic Inflammatory Response Syndrome (SIRS) criteria:

- temperature >38°C OR temperature <36°C
- respiratory rate >20 breaths/min OR $PaCO_2$<32 mmHg
- heart rate >90 beats/minute
- leucocyte count <4,000 mm^3 OR >12,000 mm^3

and organ dysfunction, hypoperfusion, or hypotension. From now on we refer to these criteria as the *Bone* criteria.

Patients who have this combination of symptoms may have sepsis, however, this is not necessarily the case. We refer to these patients as being *possibly septic*. On the other hand, we can define a group of patients that are septic for sure, namely those who fulfill the Bone criteria and have severe multiple organ failure. We will refer to these patients as the *definitely* septic patients and define them as fulfilling the *strict* criteria: the Bone criteria plus at least three of the following symptoms of organ failure:

- pH $\leq$ 7.30
- thrombocyte count $<$ 80,000 mm^3
- urine output $<$ 0.5 ml/kg body weight/hour (provided the patient is not on chronic dialysis),
- $PaO_2/FiO_2 \leq 250$, and
- systolic blood pressure <90 mmHg OR vaso-active medication.

3 Rough DL for vague knowledge

We now present a conservative extension of Description Logics (DLs), i.e. an extension which improves the modeling capacities without changing the expressive power of the language. More concretely, we will introduce two modal-like operators $(\underline{\cdot})$ and $(\overline{\cdot})$ for lower and upper approximations to describe elements which either belong definitively or possibly to the concepts under its scope. These operators introduce a notion of approximation without effectively increasing the expressiveness of the language. Thus, we get extra modeling facilities for free, without having to develop new calculi, and without paying an extra price in computational complexity.

3.1 Description Logics

Description Logics (DL) are a well-studied family of set-description languages which usually come with (some or all) Boolean operators and limited quantification, and which can be extended with additional functionality in a modular way. This way properties on relations (such as symmetry, transitivity or inclusion hierarchies), number restrictions, or even some form of data-types (Concrete Domains) are often included. Description Logics have a well-defined model-theoretic semantics, and the last two decades the computational properties of a wide variety of DLs has been studied.

Formally, we introduce the DL $\mathcal{ALC}$, which is sufficient to model our case-study. The general definition of approximations, however, will be independent of any particular DL. $\mathcal{ALC}$ is a simple DL with conjunction $C \sqcap D$, disjunction $C \sqcup D$, negation $\neg C$ and universal $\forall r.C$ and existential quantification $\exists r.C$. The semantics is given as follows:

Def. 1 *Let $\mathcal{I} = (U, \cdot^{\mathcal{I}})$ be an interpretation, where U is a universe, and $\cdot^{\mathcal{I}}$ a function mapping concept names to subsets and role names to relations over U. It extends to the Boolean operators as usual and to the quantifier as follows:*

- $(\exists R.C)^{\mathcal{I}} = \{i \in U \mid \exists j \in U : (i,j) \in R^{\mathcal{I}} \;\&\; j \in C^{\mathcal{I}}\}$
- $(\forall R.C)^{\mathcal{I}} = \{i \in U \mid \forall j \in U : (i,j) \in R^{\mathcal{I}} \rightarrow j \in C^{\mathcal{I}}\}$

In a terminology $\mathcal{T}$ (called *TBox*) the interpretations of concepts can be restricted to the *models* of $\mathcal{T}$ by *axioms* of the form $C \sqsubseteq D$ or $C \doteq D$. Based on this model-theoretic semantics, concepts can be checked for *unsatisfiability*: whether they are necessarily interpreted as the empty set. Another useful semantic implication is *subsumption* of two concepts C and D (a subset relation of $C^{\mathcal{I}}$ and $D^{\mathcal{I}}$ w.r.t. all models $\mathcal{I}$ of $\mathcal{T}$) denoted by $\mathcal{T} \models C \sqsubseteq D$.

A *knowledge base* $\Sigma = (\mathcal{T}, \mathcal{A})$ extends a TBox $\mathcal{T}$ with an assertional component (usually called *ABox*) $\mathcal{A}$, which is a set of assertions $i : C$ and $R(i, j)$ for individual names i, j, a relation R and a concept C. The semantics is a straightforward extention of the previous definition: an interpretation $\mathcal{I}$ is a model for a assertions $i : C$ and $R(i, j)$ if, and only, $i^{\mathcal{I}} \in C^{\mathcal{I}}$ and $R^{\mathcal{I}}(i^{\mathcal{I}}, j^{\mathcal{I}})$. Then, a knowledge base is *consistent*, if there is a model for both its TBox and ABox.

3.2 Rough Description Logics

Description Logics are suitable for modeling crisp knowledge but are often too rigid for approximate information. For example, no explicit mechanism is in place when a definition is not commonly agreed upon, or when exceptions need to be captured. The sepsis use-case provides an example for such vaguely defined classes, for which *no agreed upon criteria* exist to determine whether a patient is indeed septic or not.

The basic idea is rather straightforward: even though we fail to formally define the class of septic patients, we can approximate it by giving an upper and a lower bound. The *upper approximation* of the set of septic patients is formed by the set of patients that fulfill the Bone criteria, i.e. the **possibly** septic patients. Orthogonally, the *lower approximation* of the set of septic patients is the set of patients that are **definitely** septic, i.e. the patients that fulfill the strict criteria.

Traditionally, in DLs this is modeled using *primitive* definitions, i.e. axioms of the form $C \sqsubseteq D$, where C is restricted by D without being fully defined. The relation between the concept *Septic* and its approximations is in the pure DL modeling just ***Definitely*_*Septic*** $\sqsubseteq$ *Septic* $\sqsubseteq$ ***Possibly*_*Septic***.

Rough DL: Approximations, Syntax and Semantics Modelling vague concepts with the traditional approach has its limits when the vague concept of *Septic* patients needs to be defined. Let us consider a special type of sepsis where the renal system fails. In DL terms, the relation between renal sepsis and sepsis would be modeled by an axiom *RenalSeptic* $\sqsubseteq$ *Septic*. Again, renal sepsis is not definable in a crisp way, but there could be an approximation describing patients which have possibly renal sepsis. Now, the question arises whether possibly renal septic patients should be possibly septic, i.e. whether ***Possibly*_*RenalSeptic*** $\sqsubseteq$ ***Possibly*_*Septic*** or not. In traditional DL it is possible to have all typical properties of the renal sepsis, but not the typical properties of a sepsis. What is missing is automatic inheritance of the approximations in a monotonic way.

$$\begin{tikzcd}
\textit{Possibly\_RenalSeptic} \arrow[r, dash, "???"] & \textit{Possibly\_Septic} \\
\textit{RenalSeptic} \arrow[u, dash, "\sqsubseteq"] \arrow[r, dash, "\sqsubseteq"] & \textit{Septic} \arrow[u, dash, "\sqsubseteq"] \\
\textit{Definitely\_RenalSeptic} \arrow[u, dash, "\sqsubseteq"] \arrow[r, dash, "???"] & \textit{Definitely\_Septic} \arrow[u, dash, "\sqsubseteq"]
\end{tikzcd}$$

In our motivating picture there should be subsumption relations at the "???" positions, i.e. that ***Definitely*_*RenalSeptic*** $\sqsubseteq$ ***Definitely*_*Septic*** and ***Possibly*_*RenalSeptic*** $\sqsubseteq$ ***Possibly*_*Septic*** should be a logical consequence of the knowledge base. In this sense, DL is inappropriate to model vague information, as there is a stronger semantic relations underlying the approximations of a concept. With Rough Description Logics (Rough DL), which we are about to introduce, we attempt to close this gap in a conceptually simple way.

Before providing formal semantics it is worth pointing out that approximations have very distinct properties. The upper approximation is the set of patients with a strong indication that they might be septic. Formally, this means that for every patient pat_1 in ***Possibly*_*Septic***, there must be at least one septic patient pat_2, for which there are no criteria to explain why pat_2 differs from pat_1, i.e. pat_1 is indiscernible from pat_2.

Rough DL is not restricted to a particular DL, and will be defined for an arbitrary Description Logic $\mathcal{DL}$.

Def. 2 *The language $\mathcal{RDL}$ of Rough DL is the smallest set of concepts containing $\mathcal{DL}$, and for every concept $C \in \mathcal{RDL}$ also the* upper approximation $\overline{C} \in \mathcal{RDL}$ *and the* lower approximation $\underline{C} \in \mathcal{RDL}$.

The notions of rough T- and ABox, as well as rough knowledge base extend the usual notions in the expected way.

Def. 3 *Let a rough interpretation be a triple $\mathcal{I} = (U, R^{\sim}, \cdot^{\mathcal{I}})$, where U is a universe, $\cdot^{\mathcal{I}}$ an interpretation function, and $R^{\sim}$ an equivalence relation over U. The function $\cdot^{\mathcal{I}}$ maps $\mathcal{RDL}$ concepts to subsets and role names to relations over the domain U. It extends to the new constructs as follows:*

- $(\overline{C})^{\mathcal{I}} = \{i \in U \mid \exists j \in U : (i, j) \in R^{\sim} \;\&\; j \in C^{\mathcal{I}}\}$
- $(\underline{C})^{\mathcal{I}} = \{i \in U \mid \forall j \in U : (i, j) \in R^{\sim} \rightarrow j \in C^{\mathcal{I}}\}$

Intuitively, the upper approximation of a concept C covers the elements of a domain with the *typical properties* of C, whereas the lower approximation contains the *prototypical* elements of C.

What did we gain? Even if it is impossible to formally define a concept, such as *Septic*, we can often specify the approximations. In our use-case, the upper approximation can be defined using the "Bone criteria", the lower approximation, using the set of "Strict criteria" described in Section 2. In Rough DL we now model vague knowledge in a precise way; with explicit formal semantics.

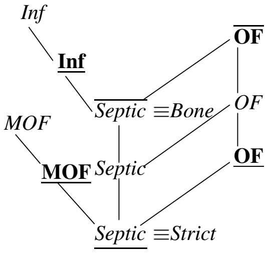
Some logical consequences of the semantics Consider a simplistic Rough DL terminology, which models sepsis by its approximations. Concretely, having an infection is a certain property of **possibly septic** patients , i.e. the upper approximation $\overline{\textit{Septic}}$ is a subconcept of *Inf.* Also, septic patients must have an organ failure (OF) in at least one organ system. Furthermore, **definitely septic patients** must have multiple organ failure. This gives the following terminology: $\mathcal{T}$ $=\{\overline{\textit{Septic}} \doteq$ Bone, $\underline{\textit{Septic}} \doteq$ Strict, $\overline{\textit{Septic}}$ $\sqsubseteq$Inf, Septic$\sqsubseteq$ OF, $\underline{\textit{Septic}}$ $\sqsubseteq$MOF$\}$ With the implicit semantics of $\mathcal{RDL}$ there are logical consequences, some of which we will discuss in more detail.

- *Possibly septic patients must be definitely infected.* In logical terms, we have $\mathcal{T} \models \overline{\textit{Septic}} \sqsubseteq \underline{\textit{Inf}}$. Why is this the case? Take a patient pat with all the typical properties of sepsis, including an infection. Assume that he has an atypical infection, i.e., that there is a similar patient

pat_2 without an infection. But pat being typically septic means that there must be a septic patient pat_3 similar to pat, to which pat_2 is also similar because of transitivity. Then pat_2 is similar to a septic patient, and must belong to the upper approximation $\overline{\mathit{Septic}}$. By this he must have all the typical properties of sepsis, including an infection, which is a contradiction.

- *Possibly septic patients must have possible organ failure.* Formally, we can conclude that $\mathcal{T} \models \overline{\mathit{Septic}} \sqsubseteq \overline{\mathit{OF}}$. This means that if we know that organ failure is part of the proper definition of sepsis, patients that are possibly septic must at least have some condition that resembles organ failure. A similar result holds for the lower approximation $\underline{\mathit{Septic}}$ and multiple organ failure.

  The following figure shows the taxonomy of axioms based on the subsumption hierarchy w.r.t. the rough DL semantics, where the relations with the boldly printed concepts are implicitly derived.



Inf
$\underline{\text{Inf}}$
$\overline{\textbf{OF}}$
$\overline{\mathit{Septic}}$ ≡*Bone*
OF
MOF
$\underline{\textbf{MOF}}$
Septic
$\underline{\textbf{OF}}$
$\underline{\mathit{Septic}}$ ≡*Strict*

There are more examples of the intrinsic semantics of Rough DLs, which do not show in the previous figure.

- *There are no definitely non-typical sepsis patients.* Suppose that we define non-typical sepsis patients (*NTS*) as those septic patients which are not definitely septic, i.e., patients for which a similar patient exists which is not diagnosed as septic. Formally, we add an axiom $\mathit{NTS} \sqsubseteq \mathit{Septic} \sqcap \neg \underline{\mathit{Sepsis}}$ to $\mathcal{T}$ to get a new TBox $\mathcal{T}'$. Rough DL semantics implies that there can be no definitely non-typical septic patients, i.e. that $\mathcal{T}' \models \underline{\mathit{NTS}} = \bot$.
- *Definitively septic or definitively not septic.* Suppose that for a new trial only patients are selected which are definitively only diagnosed as either definitively septic, or definitively not septic, i.e., $\underline{\forall\, \mathit{diag}.(\underline{\mathit{Septic}} \sqcup \underline{\neg \mathit{Septic}})}$. Then, every patient who **is** diagnosed as possibly septic, $\exists \mathit{diag}.\overline{\mathit{Septic}}$, must possibly have been diagnosed as definitively septic (or $\overline{\exists\, \mathit{diag}.\underline{\mathit{Septic}}}$).
- Finally, it is a simple consequence of the semantics that approximations of approximations are equivalent to the approximations themselves, e.g., that $\underline{\underline{\mathit{Septic}}} \equiv \underline{\mathit{Septic}}$.

Reasoning with Rough DLs One of the main advantages of our newly introduced modeling mechanism is that reasoning almost comes for free. As opposed to most other mechanisms to deal with vague knowledge in DL, reasoning with approximations can be reduced to standard DL reasoning, by translating rough concepts into pure DL concepts with a special reflexive, transitive and symmetric role.

Let C be a rough concept. We define a translation function $(\cdot)^t : \mathcal{RDL} \rightarrow \mathcal{DL}$ for concepts with $A^t = A$ for atomic concepts A, and $(\overline{C})^t = \exists r^{\sim}.C$, and $(\underline{C})^t = \forall r^{\sim}.C$ for $C \in \mathcal{RDL}$ where $r^{\sim}$ is a new role symbol, and where the translation function is inductively applied on subconcepts for all other constructs. This definition can be extended to axioms $(C \sqsubseteq D)^t = C^t \sqsubseteq D^t$ and TBoxes $\mathcal{T} = \{ax_1, \ldots, ax_n\}$ as follows: $\mathcal{T}^t = \{refl(r^{\sim}), sym(r^{\sim}), trans(r^{\sim}), ax_1^t, \ldots, ax_n^t\}$.

For any DL $\mathcal{DL}$ with universal and existential quantification, and symmetric, transitive and reflexive roles, there is no increase in expressive power, i.e. Rough DL can be simulated in (almost) standard DL.

Prop. 1 *Let $\mathcal{RDL}$ be the rough extension of a Description Logic $\mathcal{DL}$, $\mathcal{T}$ an $\mathcal{RDL}$ TBox, and $(\cdot)^t$ the above given translation. An $\mathcal{RDL}$ concept C is satisfiable in a rough interpretation w.r.t. $\mathcal{T}$ iff C is $\mathcal{DL}$-satisfiable w.r.t. $\mathcal{T}^t$. Formally: $\mathcal{T} \models C = \bot$ iff $\mathcal{T}^t \models C^t = \bot$.*

The proof is by contradiction: assume that $\mathcal{T} \models C = \bot$ in Rough DL, but that there is a DL model $\mathcal{I} = (U, (\cdot)^{\mathcal{I}})$ of $\mathcal{T}^t$ such that $(C^t)^{\mathcal{I}} \neq \varnothing$. It follows from the construction of the translation function $(\cdot)^t$ that $\mathcal{I}' = (U, r^{\sim}, (\cdot)^{\mathcal{I}'})$ is a model for $\mathcal{T}$, and that $C^{\mathcal{I}'} \neq \varnothing$, which is a contradiction. The other direction is similar.

As with usual DLs, one can reduce other reasoning services, such as subsumption, to satisfiability (and finally to ABox consistency) in the presence of negation. Rough DL are no different. As the translation is linear, the complexity of reasoning in Rough DL is the same as of reasoning in its carrier DL with quantifiers, symmetry and transitivity.

4 Modeling Clinical trials with Rough DL

Clinical trials use *entry criteria* to select patients for the study. The choice of these criteria is an important step in clinical trial design: to be able to compare the results of the trial with those of other trials and to assess the generalizability of the results to daily clinical practice, the entry criteria have to be compatible with definitions used in comparable trials and the agreed standard definitions of disease. This is obviously complicated when no crisp disease definition exists.

In the case of severe sepsis, nine recent randomized clinical trials all used different entry criteria to select patients with severe sepsis [Peelen *et al.*, 2005]. Seven out of the nine investigated trials used a structure similar to the original consensus definition for severe sepsis: confirmed infection plus SIRS criteria plus organ failure. However, the number of required SIRS criteria varied between the trials and some trials used a slight modification of the original SIRS criteria. Furthermore, the specification of organ failure and the required number of failing organ systems differed.

One way to investigate the differences in entry criteria is to compare the definitions used in the trials with the approximations of the medical condition. In our study, we use the concepts *Strict* and *Bone* as approximations of sepsis and compare them to the entry criteria used in the nine trials. There are four interesting situations. Are there patients that are

1. in one of the trials but not in *Bone*?

2. in all trials but not in *Strict*?
3. in *Bone* but not in one of the trials?
4. in *Strict* but not in all trials?

The existence of such patients would signal a discrepancy between the trial definitions and the interpretation of sepsis, pointing to potential flaws in the set-up of the trials. With $\mathcal{RDL}$ a validation of these flaws comes for free as it allows the user to model their assumptions about the inherent vagueness of the definitions in a precise way. We will now describe how we used $\mathcal{RDL}$ to perform such an investigation.

In order to use $\mathcal{RDL}$ for patient selection we first translated the definition for each trial into a DL formula. We did the same for the *Bone* definition and the *Strict* definition of sepsis, thus building a TBox with 11 definitions for septic patients. In addition we have translated a dataset from the Dutch National Intensive Care Evaluation (NICE) registry containing information on 71,929 patients into an ABox, using the terminology from the TBox. With the selection criteria for the different trials and the translated data, we used a DL-reasoner (Racer [Haarslev and Möller, 2001]) to select the patients that would be eligible for the different trials (thereby mimicking the patient selection process). The following table shows the numbers of patients of 4 of the 9 trials:

Definition	# patients		Definition	# patients
BONE-sepsis	5633		Lexipafant-sepsis	1607
Strict	982		OPTIMIST-sepsis	5088
UnionOfTrials	6203		PROWESS-sepsis	6201
IntersectionOfTrials	534		2SPLA2I-sepsis	4002

To answer the aforementioned questions, we define $\overline{Sepsis} \equiv Bone$ and $\underline{Sepsis} \equiv Strict$ in our $\mathcal{RDL}$ terminology, as those are the most widely accepted upper and lower approximations of *Sepsis*. Additionally, we can model the relation of the trials to the concept *Sepsis* explicitly. Although the 9 different trials widely cover different ways of describing possibly septic patients, it might be conceivable that there are patients outside the scope of all of these trials. However, one could assume that the 9 trials cover the *most typical* of *all possible* sepsis patients. Because $\mathcal{RDL}$ provides formal representations for the intuitions 'most typical' and 'all possible', we can model this assumption in a formal way. Namely, the union of all trials is equivalent to the lower approximation (i.e. the typical cases) of the upper approximation (i.e. all possible cases) of *Sepsis*. Similarly, we can model the assumption that the intersection of all trials covers the *most typical* patients that are *definitively septic*. This is done by defining the intersection of all trial concepts to be a lower approximation (i.e. the most typical cases) of the lower approximation (i.e. the definitively septic patients) of the concept *Sepsis*.

Given our experimental setup it is easy to show that there are serious flaws in the trial selection. It is a consequence of the semantics of $\mathcal{RDL}$ that an approximation itself can not be approximated. This implies that *Bone* $\equiv$ *UnionOfTrials* and *Strict* $\equiv$ *IntersectionOfTrials*. This resulted in inconsistency of the definitions with respect to the trial data.

Using our infrastructure one can now perform a more detailed *data-based validation* to detect the source of the logical contradiction. For example, we queried for patients with queries like $\neg$*Bone* $\sqcap$ *trial-X* to look for violations of the upper approximation and queries like *Strict* $\sqcap$ $\neg$*trial-X* for violations of the lower approximation. In this way, we found 141 patients in *PROWESS-sepsis* and 6 patients in *Lexipafant-sepsis* that do not fulfill the *Bone* criteria.

Finally, we can use purely *terminological reasoning* to analyse the trial criteria. For example, classifying all definitions brought to light that none of the concepts describing the trials is subsumed by *Bone*. This is an interesting result when compared to the data-based validation. Although for 7 of the trial definitions we did not find any patient that violated the upper approximation, such patients can exist in principle. Similarly, with respect to the lower approximation, we found that only 4 of the trial definitions subsumed *Strict*.

Advantage over standard DL Trial validation using a standard DL infrastructure without the rough extension is already a significant improvement over the current situation, in which patient selection is procedurally performed as a sequence of database queries. Using standard DL we can check violations, as discussed above, with A-box reasoning over the data set and the terminology, or purely terminologically, as suggested in the previous two paragraphs (which are not necessarily restricted to $\mathcal{RDL}$).

Modeling the definitions in $\mathcal{RDL}$ gives an additional improvement: the validation against the criteria is done automatically. There is a way of achieving the same validation with pure DL, which we is much less elegant, though. Here, one would sequentially check the validation criteria 1 to 4 introduced above, i.e. by checking satisfiability of the concept $\neg$*Bone* $\sqcap$ *trial-X* for all trials. However, this amounts to a procedural verification of the assumptions about the vague definitions about *Sepsis*, which is error-prone, and tedious.

Moreover, our $\mathcal{RDL}$ model excludes the invalid definitions automatically. For example, the set of patients in *Lexipafant-sepsis* $\sqcap$ $\neg$*Bone* is empty per definition. To achieve the same result in a pure DL TBox one has to model the relation between trials and *Bone* explicitly, e.g. by asserting *UnionOfTrials* $\equiv$ *Bone*. But this is an incorrect oversimplification of the relation between the trials and the approximation of *Sepsis* as opposed to the much more accurate $\mathcal{RDL}$ formalisation.

5 Related and future work

The work described in this paper covers a wide variety of topics that have been studied extensively in the literature. This means that there are a plethora of similar approaches, to which we will briefly refer.

- *Rough DL versus Modal DL.* From a technical perspective, Rough DL is a *fusion* of DL with modal **S5**.[1] Most attempts to introduce modal operators into DL focus on unions or produces, which usually requires more complex, mostly Kripke-based, semantics (e.g., [Baader and Laux, 1995]) and new decision procedures. Our modalities range over the domain itself rather than over varying

[1] Fusion means, that the different operators of the two languages apply on different sets of roles, and don't interfere [Baader *et al.*, 2002]. This makes fusions behave better than their more complex relatives unions or products [Gabbay *et al.*, 2003].

domains, which makes them easier to handle, e.g., decidability and complexity results come for free, and we can apply existing reasoners.

- *Rough DL versus Fuzzy DL.* Fuzzy DLs have recently got increasing attention, particularly starting with the work of Straccia [Straccia, 2001]. Vagueness of concepts is expressed as a degree of membership. Rough DL advocate a simpler and *qualitative* approach, which is appropriate for some domains, such as the medical. In our case study, e.g., there is no way of quantifying membership of the class *Septic*, but well-defined upper and lower approximations. Note, for example, that the Bone criteria are define in a crisp, non-fuzzy way.
- *Rough DL versus Rough Sets.* The connection between Rough Set theory and modal logic is well-established [Düntsch, 1997], and there have been previous attempts to introduce concept languages to model approximations [Orlowska, 1988]. Orlowska's work, which is closed to our own, is restricted to propositional logic and is, to the best of our knowledge, neither implemented nor practically applied and evaluated. An interesting orthogonal approach in [Doherty *et al.*, 2003] where concepts are defined as pairs of approximations. However, their semantics is non-standard, and approximate concepts cannot as easily be integrated in standard ontology languages as with Rough DL.
- *Rough DL versus Defaults.* Rough DL can also be useful to model defaults as one can use lower approximations to capture exceptions in an intuitive way. Simply speaking, the lower approximation then contains the *typical* subset of the elements of a concept. Further discussion of this idea is out of the scope of this paper.

Extension and alternative definitions The restriction of the semantics to equivalence relations goes back to Pawlak's work [Pawlak, 1982]. To model vague concepts, one might also study approximation operators based on tolerance relations (reflective and symmetric). Also one could think of sets of equivalence classes according to different similarity relations. An interesting extension to graded rough modalities, as suggested in [Yao and Lin, 1996] is easily integrated into Rough DL, as they can be translated into number restrictions.

Before extending the language, the more pressing issue of efficiency of the reasoning has to be solved. So is Racer, our current DL reasoner, not optimised for reasoning with equivalence classes, which makes reasoning sometimes inefficient.

A different path for future research is the explicit integration of equivalence relations into Rough DL ABoxes. Often, data can be classified into indiscernible clusters. In a first step, Rough DL can be a suitable query language, but it is also conceivable to learn Rough DL concepts from the explicit definitions of the instances of particular concepts.

6 Conclusions

Rough DL, the extension to standard DLs that we introduce in this paper, allows for precise modeling of vague knowledge. Modeling vague knowledge is a common need in realistic domains, e.g. in medicine. An advantage of modeling concept approximations in a qualitative way is that queries to the model give crisp answers. We have shown that reasoning in $\mathcal{RDL}$ can be reduced to standard DL satisfiability, which gives us access to reasoning infrastructure.

In our evaluation of medical trials about sepsis patients we have shown that modeling vague knowledge can help to answer important questions in the design of clinical trials. The validation of trials based on their formal definitions is an improvement over the usual data-based validation. When the validation declaratively is done using Rough DL, the logical consequences of the semantics immediately reveals inconsistencies in the trial definitions, whereas several successive queries are necessary to do the same with standard DLs. Finally, we claim that Rough DL can be very useful when building new trials with vaguely defined medical conditions, as they enforce better models for the selection of patients.

References

[Baader and Laux, 1995] F. Baader and A. Laux. Terminological logics with modal operators. In *Proc. of IJCAI*, pages 808–814, 1995.

[Baader *et al.*, 2002] F. Baader, C. Lutz, H. Sturm, and F. Wolter. Fusions of description logics and abstract description systems. *JAIR*, 16:1–58, 2002.

[Baader *et al.*, 2003] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. Patel-Schneider, editors. *The DL Handbook*. Cambridge University Press, 2003.

[Bone, R.C., 1992] Bone, R.C. Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. *Crit Care Med*, 20(6):864–874, 1992.

[Doherty *et al.*, 2003] P. Doherty, M. Grabowski, W. Lukaszewicz, and A. Szalas. Towards a framework for approximate ontologies. *Fundam. Inf.*, 57:147–165, 2003.

[Düntsch, 1997] I. Düntsch. A logic for rough sets. *Theoretical Computer Science*, 179(1-2):427–436, 1997.

[Gabbay *et al.*, 2003] D.M. Gabbay, A. Kurucz, F. Wolter, and M. Zakharyatschev. *Many-Dimensional Modal Logic: Theory and Applications*. Elsevier, 2003.

[Haarslev and Möller, 2001] V. Haarslev and R. Möller. RACER system description. In R. Goré, A. Leitsch, and T. Nipkow, editors, *IJCAR*, number 2083 in LNAI, 2001.

[Orlowska, 1988] E. Orlowska. Logical aspects of learning concepts. *Int. J. of Approx. Reasoning*, 2:349–364, 1988.

[Pawlak, 1982] Z. Pawlak. Rough sets. *Int. J. of Computer and Information Sciences*, 11:341–356, 1982.

[Peelen *et al.*, 2005] L. Peelen, N.F. De Keizer, N. Peek, E. De Jonge, R.J Bosman, and G.J. Scheffer. Influence of entry criteria on mortality risk and number of eligible patients in recent studies on severe sepsis. *Crit Care Med*, 33(10):2178–2183, 2005.

[Straccia, 2001] Umberto Straccia. Reasoning with fuzzy description logics. *J. of AI Research*, 14:137–166, 2001.

[Yao and Lin, 1996] Y.Y. Yao and T.Y. Lin. Generalization of rough sets using modal logics. *Intelligent Automation and Soft Computing*, 2(2):103–120, 1996.