

---

Aleksander Øhrn

---

# Discernibility and Rough Sets in Medicine: Tools and Applications

---

Extended Abstract

Department of Computer and Information Science  
Norwegian University of Science and Technology  
N-7491 Trondheim, Norway



## Summary

This thesis examines how discernibility-based methods can be equipped to possess several qualities that are needed for analyzing tabular medical data, and how these models can be evaluated according to current standard measures used in the health sciences. To this end, tools have been developed that make this possible, and some novel medical applications have been devised in which the tools are put to use.

Rough set theory provides a framework in which discernibility-based methods can be formulated and interpreted, and also forms an appealing foundation for data mining and knowledge discovery. When the medical domain is targeted, several factors become important. This thesis examines some of these factors, and holds them up to the current state-of-the-art in discernibility-based empirical modelling. Bringing together pertinent techniques, suitable adaptations of relevant theory for model construction and assessment are presented. Rough set classifiers are brought together with ROC analysis, and it is outlined how attribute costs and semantics can enter the modelling process.

ROSETTA, a comprehensive software system for conducting data analyses within the framework of rough set theory, has been developed. Under the hypothesis that the accessibility of such tools lowers the threshold for abstract ideas to migrate into concrete realization, this aids in reducing a gap between theoreticians and practitioners, and enables existing problems to be more easily attacked. The ROSETTA system boasts a set of flexible and powerful algorithms, and sets these in a user-friendly environment designed to support all phases of the discernibility-based modelling methodology. Researchers world-wide have already put the system to use in a wide variety of domains.

By and large, discernibility-based data analysis can be varied along two main axes: Which objects in the universe of discourse that we deem it necessary to discern between, and how we define that discernibility among these objects is allowed to take place. Using ROSETTA, this thesis has explored various facets of this also in three novel and distinctly different medical application areas:

- A method is proposed for identifying population subgroups for which expensive tests may be avoided, and experiments with a real-world database on a cardiological prognostic problem suggest that significant savings are possible.
- A method is proposed for anonymizing medical databases with sensitive contents via cell suppression, thus aiding to preserve patient confidentiality. As electronic medical records and medical data repositories get more common and widespread, the issue of making sensitive data anonymous becomes increasingly important.
- Very simple rule-based classifiers are employed to diagnose acute appendicitis, and their relative performance is compared to a team of experienced surgeons. The added value of certain biochemical tests is also demonstrated.

## Background

Medicine has from early on often been employed as a testbed domain for newly developed learning and reasoning techniques from computer science. Not only because the field of medicine has many applications whose solutions are important in a social context, but also because the field is notoriously difficult with a wide range of confounding factors and aspects that demand special considerations, hence forming a technically challenging area. This interplay between computer science and medicine has led to some remarkable work being accomplished in the last 30 years, but major developmental efforts and research is still needed if a significant impact on the practice of medicine is to be realized.

The directions to take for construction of the underlying computational models may differ. Traditionally, models from the field of artificial intelligence have been very knowledge-intensive in the sense that domain-specific knowledge about such things as etiology and pathophysiology have been carefully encoded by hand into the model. On the other end of the spectrum are approaches where models are attempted induced or synthesized from low-level data without relying on a priori domain knowledge. A shift in vogue towards such methods has been observed in the last decade, something which is also in line with an increasing emphasis in medicine on evidence-based practice.

Databases often grow so large that human inspection and interpretation of the data is not feasible, with a gap between data generation and data understanding as a result. Clearly, tools and techniques that can aid in extracting unknown interesting patterns buried in the data would be useful to help bridge this gap. Classical tools for database querying may be adequate if you know what to look for, but often the most interesting queries to pose cannot be formulated as straightforward lookups. To be able to deal with such advanced queries, more intelligent data analysis tools are needed. Research in computer science has in the last decades spawned a vast multitude of methods that “learn” from examples, and that can be used to extract patterns from empirical data for classification. Such techniques are increasingly being applied to medical data sets.

Rough set theory was introduced in the early 1980s as a tool for representing and reasoning about imprecise or uncertain information. Based on the notion of indiscernibility and the inability to distinguish between objects, rough set theory deals with the approximation of sets or concepts by means of binary relations, typically constructed from empirical data. As the methodology has matured, several interesting applications of the theory have surfaced, also in medicine.

This thesis concerns itself with tools for and applications of discernibility and rough sets in medicine. As such, the context in which to place this work is an intersection of several scientific areas, the perhaps two most relevant being the field of data mining and knowledge discovery, and the field of medical informatics. This work focuses on the development of tools and techniques from a certain subfield of the former area, and applying them in the latter.

## Objectives

Obviously, targeting the medical domain has an important social dimension since solutions to relevant problems in this domain may have a beneficial impact on human beings and their welfare. As such, this thesis constitutes an incremental contribution towards the overall long-term goal of improving the quality of healthcare and lowering costs. There is also a challenging technical dimension to targeting the medical domain, since the art of medicine is not an exact science in which processes are easily formalized or modelled. Furthermore, these two dimensions intertwine in a way that exerts tight constraints on the solution space.

Many aspects of what we might perceive as intelligent behavior can essentially be reduced to the ability to classify. And any classification task crucially relies on an ability to appropriately discern between objects or situations. Discernibility-based methods, and rough set theory in particular, are intuitively appealing and involve, technicalities aside, rather simple ideas. Also, their formal soundness defines a solid theoretical basis for empirical modelling.

The main objectives of this thesis are:

- To discuss and illuminate the usefulness of discernibility-based methods for data mining in the health sciences, and to demonstrate that these ideas are indeed applicable by devising relevant and novel medical applications.
- To furnish the research community with a set of flexible and powerful software tools for conducting data analyses within the framework of rough set theory, and to provide a software environment which facilitates such experimentation. Under the hypothesis that the accessibility of such tools lowers the threshold for abstract ideas to migrate into concrete realization, this aids in reducing a gap between theoreticians and practitioners, and enables existing problems to be more easily attacked.

## Results

The work in this thesis has been carried out as a combination of adaptations of relevant theoretical constructs, programming and computer simulations. Through this, the versatility of discernibility-based methods for empirical modelling in general has been demonstrated. By examining aspects of the medical domain, several issues for data-driven modelling that become particularly important in the context of medicine have been illuminated. In itself, this is valuable as an aid to raise the level of consciousness among data mining practitioners towards some of the demands that the medical domain imposes. Some of the most important identified issues have subsequently been held up to the current state-of-the-art in modelling based on the relatively young fields of discernibility and rough sets, and it has been shown how the relevant methodology can be suitably adapted and employed.

- Evaluation measures central in the health sciences have been carried over to the field of rough sets. It has been demonstrated how rough set classifiers can be evaluated through ROC analysis, how calibration can be appraised, and how set approximations can be assessed in terms of sensitivity and specificity. This seems to be novel in the context of rough sets.
- It has been outlined how the use of attribute costs can be embedded in the model construction process. By employing cost information in the reduction process, low-cost rather than low-cardinality solutions can be obtained. Costs can also be made use of for model filtering and evaluation purposes.
- It has been made clear how one by overloading the notion of discernibility can cater for, e.g., hierarchically ordered attribute values. In the medical domain, such hierarchies can be encountered in some controlled medical terminologies. Missing values can be perceived as a special case of this.

Starting from a theoretically well-founded approach, a modelling methodology has been established and an extensive toolkit for discernibility-based empirical modelling has been designed and implemented. The software system, ROSETTA, is a robust, user-friendly and powerful system for discernibility-based data mining and knowledge discovery, and has by design been accommodated with the necessary features for analyzing tabular data from the medical domain. The system also exhibits originality in that it is designed to cover all stages of the knowledge discovery process, and not just isolated procedures.

Although equipped with several features that are relevant for analysis of medical data, ROSETTA is in itself a general-purpose system that is not geared towards any particular application domain. ROSETTA has been put to use by a large number of researchers world-wide, and has resulted in scientific publications in a wide variety of areas. As of February 17, 2001, more than 2200 unique users have downloaded ROSETTA.

Novel medical applications have been devised, which, although diverse in theme and scope, all share a common discernibility-based foundation. The ROSETTA system has been employed for all examples and experiments throughout.

- By observing the flux in approximation regions when our indiscernibility relation changes due to the removal of some information, an approach to identification of interesting population subgroups was presented. Geared towards pinpointing objects for whom a certain battery of expensive medical tests could be avoided, a prognostic problem in cardiology was used as a case study. For the purpose of predicting future hard cardiac events, a group of patients was identified for whom a scintigraphic scan could be avoided wrt. their discernibility status. Furthermore, meta-approximations showed that this group could be fairly well circumscribed. On a general level, the proposed method is a contribution towards lowering costs and increasing efficiency in healthcare.

- A method based on Boolean reasoning is proposed for anonymizing medical databases with sensitive contents via cell suppression, thus aiding to preserve patient confidentiality. By using the prime implicants of a certain Boolean function, algorithms are proposed that introduce indiscernibility into a system in a controlled manner. In a medical setting, this is of particular interest due to privacy issues and to prevent the possible misuse of confidential information. In an increasingly computerized world, the issue of preserving confidentiality is not likely to become any less important. By showing how discernibility can provide a foundation for cell suppression, a contribution has been made to the sound management of sensitive medical data.
- Through reducing the set of discerning attributes to singletons, it was investigated how extremely simple classification rules could be used to diagnose acute appendicitis. From a clinical viewpoint, it is of interest to determine how well computer models compare with medical doctors. Extensive simulations showed that a simple computer model was on par with a team of experienced surgeons with respect to diagnosing acute appendicitis, and the added value of certain biochemical attributes was also demonstrated. Development of simple models are important if they are to migrate into clinical practice as rules of thumb.