Solving the problem of incomplete data in medical diagnosis via interval modeling

Andrzej Wójtowicz, Patryk Żywica, Anna Stachowiak,
Krzysztof Dyczkowski*

Department of Imprecise Information Processing Methods
Faculty of Mathematics and Computer Science, Adam Mickiewicz University
Umultowska 87, 61-614 Poznań, Poland

Abstract

This paper presents an approach to making accurate and high-quality decisions under incomplete information. Our comprehensive approach includes interval modeling of incomplete data, uncertaintification of classical models and aggregation of incomplete results. We conducted a thorough evaluation of our approach using medical data for ovarian tumor diagnosis, where the problem of missing data is commonly encountered. The results confirmed that methods based on interval modeling and aggregation make it possible to reduce the negative impact of lack of data and lead to meaningful and accurate decisions. A diagnostic model developed in this way proved better than classical diagnostic models for ovarian tumor. Additionally, a framework in R that implements our method was created and is available for reproduction of our results. The proposed approach has been incorporated into a real-life diagnosis support system – OvaExpert.

Keywords: missing data, uncertainty, aggregation, medical diagnosis, decision-making

1. Introduction

The aspect of data uncertainty is studied intensively in many contexts and scientific disciplines, including medicine. Many different forms of uncertainty

[*] Corresponding author
Email address: chris@amu.edu.pl (Krzysztof Dyczkowski)

in data have been recognized: one comes from conflicting or incomplete information, as well as from multiple interpretation of some phenomenon; another arises from lack of well-defined distinctions or from imprecise boundaries. Functioning under uncertainty and ignorance is an everyday experience of many practitioners, and is impossible to eliminate completely. For example, in medical practice it has been shown [1, 2] that the collection of complete data by a physician during examinations can be highly problematic due to the technical limitations of the healthcare institution, the high costs of a medical examination or the high risk of deterioration in a patient's health after a potential examination. The lack of data hinders the use of traditional models for diagnosis support, and there is therefore an urgent need to solve this problem.

One of the possible approaches to managing incompleteness of data is to exploit well-established methods from the field of data imputation (see [3]). Undoubtedly in many research areas such an approach is sufficient. However in medical applications, where human life is at stake, it is not so clear whether we can introduce new data which may be subject to small but significant error. Another option is to develop a new model specially dedicated to incomplete data. However, the multiplicity of already existing models makes it difficult to select the right one among them, and consequently physicians are confused and refrain from using any of them. Adding yet another diagnostic model would increase complexity in modeling and computation even further. For these reasons we explore an entirely different path. The main idea is to construct a general method that makes it possible to adapt and integrate existing and well-established diagnostic methods to make them usable with incomplete data.

A direct motivation of our work was the need to support gynecologists in diagnosis of ovarian tumor, including in the case of incomplete data. This type of cancer is particularly difficult to diagnose, and its mortality rates have remained high for many years ([4]). The main problem is to determine whether a tumor is malignant or benign based on two groups of parameters: data from medical history (e.g. age, weight, number of pregnancies) and diagnostic data (e.g. blood markers, ultrasonography). The research problem therefore boils down to a binary classification problem.

There are several well-known models in ovarian tumor diagnostics. Some of them are created by individual research units, such as the Alcazar model and SM; others by organizations (incorporating a number of research centers), such as IOTA LR1. The majority are scoring models and models based on logistic regression. These models attain different levels of effectiveness [5, 6], generally high on internal data but very often much lower during external validation. Different models use different patient attributes, and collecting all of them may be costly and problematic. Moreover, these models are not prepared for the case where some of the data in the patient description are missing. Recently, IOTA developed the first model that is able to handle missing value of one attribute [7]. In this paper we want to propose general method for handling missing values. The importance of the completeness and quality of medical data was recently highlighted in [8].

As a result, the ability to diagnose – called *diagnosability* or *decisiveness* – of these models may be low in many practical situations. So far, all research in this field were made on complete data sets. In consequence the problem of data incompleteness is not well investigated although it is currently discussed in medical community [9]. Furthermore, unlike in other classification problems, there is no clearly defined and widely accepted indicator of the quality of such a diagnostic model. The most commonly used metrics are the area under the ROC curve (AUC), accuracy, and sensitivity. However, these do not reflect all of the aspects of the problem analyzed here; in particular they do not take into account the level of diagnosability.

In our approach we were able to turn some of the above-mentioned drawbacks into assets, so as to achieve a solution to the problem of incomplete data. During the research we noticed that, since the models use different attributes, they complement one another, allowing better decisions to be made. However, gynecologists were not yet able to take advantage of this fact. To change this, we developed the idea of creating a decision support system that would integrate knowledge derived from a number of models, and provide it in an accessible way to the doctor.

We have developed OvaExpert, a specialist diagnostic system to support gynecologists, including those less experienced, in the proper differentiation of tumors. The results presented in this paper answer problems encountered during work on the OvaExpert system. The system is currently being intensively tested at a number of medical centers. The main objective of the system is to make accurate decisions despite a lack of data. This is achieved by interval modeling of incomplete information. The use of the diversity of diagnostic models allows us to increase the efficiency of diagnosis by aggregating knowledge from many sources. To this end, we implemented a number of aggregation operators and conducted a set of tests to verify how those operators act on real-life data, both complete and incomplete. By sharing our work through GitHub, we enable other researchers to verify our results and to reuse our code for their own purposes. We believe that our results may prove valuable not only in ovarian tumor diagnosis, but also in other classification tasks in which the problem of missing and incomplete data is faced.

The remainder of the paper is organized as follows. In Section 2 we present details of our approach to dealing with data incompleteness, including uncertaintification of patient descriptions and diagnostic models, and methods of information aggregation. Section 3 describes an evaluated dataset as well as the evaluation procedure. In Section 4 the results of our experiments are presented and discussed. Section 5 emphasizes the significance of our results by giving a short introduction to their application in the OvaExpert system. Conclusions appear in Section 6.

2. Proposed approach

The main objective of our approach is to enable effective decision-making, in spite of missing data. The most obvious approach, based on imputation, is not feasible here for many reasons. First of all we were limited by the very small number of cases that could be used as a prior knowledge for effective imputation. Besides, as has already been mentioned in the previous section, imputing the results of diagnostic tests, even though it may be correct from a statistical point of view, can lead to significant diagnostic error. Imputation can serve as a convenient way of carrying out statistical analyses of a dataset or a classifier, but it must be clearly stated that imputed data are not the real one so it may be hazardous to use them for making a diagnosis for one particular patient. This issue was widely discussed in a recent book by Hatch [9]. Our primary objective was not to make an illusion of operating on complete data. We want a doctor to be aware of the incompleteness of the knowledge about a patient's state and rather to suggest no diagnosis then the wrong one. Finally, our ultimate goal is to develop a general method that deals not only with totally incomplete (missing) data but also with data complete only to some extent (interval data), for which imputation is not the answer.

In our research, we adopted the following two assumptions. Firstly, we accept a state in which a diagnostic model does not return any diagnosis. This should not happen too often, but in the most difficult diagnostic cases (or if a significant part of attributes is missing) it may be the only option. Secondly, we do not intend to create new diagnostic models.

Instead, we enable the use of existing models under missing and incomplete data. We base our research on available regression and scoring models. Theoretical example of such model as well as our approach is illustrated in the following subsections (Examples 1–4). More details about the models are given in Subsection 3.1.

2.1. Interval modeling

In a classical approach, a patient is modeled as a vector $\mathbf{p}$ in a space P. Let $D_1, ..., D_n$ be real closed intervals denoting domains of attributes that describe patients. We define a set P in the following way $P := D_1 \times ... \times D_n$. Then, a vector $\mathbf{p}$ that describes a patient has the form $\mathbf{p} = (p_1, p_2, ..., p_n)$, where $p_i \in D_i$.

A diagnostic model can be formalized as a function $m : P \to [0, 1]$. The values returned by the function indicate confidence as regards the malignancy of a tumor, and are interpreted in the following way:

- $m(\mathbf{p}) \geq 0.5$ – diagnosis towards malignant (higher values represent higher confidence);
- $m(\mathbf{p}) < 0.5$ – diagnosis towards benign (lower values represent higher confidence).

Observe that the situation where $m(\mathbf{p}) = 0.5$ is resolved towards malignancy.

Example 1. *For the sake of simplicity, in this example we assume that the patient is described only by two attributes, namely patient's age and one cancer antigen test. We define the domains of these attributes as* $D_1 = [0, 100]$ *and* $D_2 = [0, 1500]$*. Consider the following two patients:* $\mathbf{p}^A = (35, 100)$ *and* $\mathbf{p}^B = (60, 1200)$*. Let* $m_1 : P \to [0, 1]$ *be a simple example diagnostic model defined by*

$$m_1(\mathbf{p}) = 0.0025p_1 + 0.0005p_2\,.$$

Now we can easily see that according to diagnostic model m_1 *patient* A *should be diagnosed as benign (*$m_1(\mathbf{p}^A) = 0.138$*) and patient* B *as malignant (*$m_1(\mathbf{p}^A) = 0.75$*).*

The existing diagnostic models operate on complete patient data. In order to represent missing values we have to add a special element (in practice commonly denoted by NA) to the domain of each attribute. Thus patient is now described by a vector $\mathbf{p} = (p_1, ..., p_n)$, where $p_i \in D_i \cup \{\text{NA}\}$. A major disadvantage of this approach is the need to introduce a new, separate value to represent missing values. This value cannot be handled natively by the original diagnostic models, which in turn leads to an inability to make any diagnosis. Therefore we use a different approach in which all the data are represented in the same, consistent way (cf. [10]).

For each attribute D_i we introduce its interval version, defined as the set of all nonempty closed subintervals of D_i

$$\hat{D}_i = \mathcal{I}_{D_i} = \{[a,b] : \ [a,b] \subseteq D_i\} \ .$$

Analogously as before we define $\hat{P} = \hat{D}_1 \times \hat{D}_2 \times ... \times \hat{D}_n$. Throughout the reminder of this paper we will consistently use the *hat* symbol to indicate interval values in order to distinguish them from numeric ones.

In this model a patient is described by a vector of intervals

$$\hat{\mathbf{p}} = (\hat{p}_1, ..., \hat{p}_n) = \left([\underline{p}_1, \overline{p}_1], ..., [\underline{p}_n, \overline{p}_n]\right) \ .$$

We say that vector $\mathbf{p} \in P$ is an embedded vector of $\hat{\mathbf{p}} \in \hat{P}$, denoted by $\mathbf{p} \in_E \hat{\mathbf{p}}$, when

$$\forall_{1 \le i \le n} \, p_i \in \hat{p}_i \, .$$

Consequently, for each vector $\mathbf{p} \in P$ (with or without missing values) we can define its interval equivalent $\hat{\mathbf{p}} \in \hat{P}$ in the following way:

$$\underline{p}_i = \begin{cases} p_i & \text{if } p_i \neq NA \\ \min\limits_{d \in D_i} d & \text{if } p_i = NA \end{cases}, \quad \overline{p}_i = \begin{cases} p_i & \text{if } p_i \neq NA \\ \max\limits_{d \in D_i} d & \text{if } p_i = NA \, . \end{cases} \tag{1}$$

The above definition allows one to describe the value of each attribute of a patient by an interval, regardless of whether or not the description of the attribute was given. If the value was not provided then the proposed representation has the form of a set containing all possible values for the attribute. If the value was given, it is represented by an interval reduced to a point. The main advantage of this approach is that all patients can be described in the same, uniform way and can be processed with the same diagnostic model.

2.2. Uncertaintification of prediction models

The next step is to enable the diagnostic models to work with the interval representation of the patient data. We utilize a classical method of extending real functions to interval values [11] to obtain a new, *uncertaintified* diagnostic model $\hat{m}$ defined as

$$\hat{m}(\hat{\mathbf{p}}) = \{m(\mathbf{p}) : \mathbf{p} \in_E \hat{\mathbf{p}}\} \ . \tag{2}$$

The resultant interval represents all of the possible diagnoses that can be made based on a patient description in which every missing value has been replaced with all possible values for that attribute. The more incomplete the description, the more uncertain the diagnosis. However, it is worth noting that in many cases it is still possible to make a proper decision, since some amount of missing values is acceptable and will not affect the final result significantly.

One would expect that the result of reasoning based on an interval representation will also be an interval. The value of $\hat{m}(\hat{\mathbf{p}})$ can also be defined in interval form

$$\hat{m}(\hat{\mathbf{p}}) = \left[\min_{\mathbf{p} \in_E \hat{\mathbf{p}}} m(\mathbf{p}), \max_{\mathbf{p} \in_E \hat{\mathbf{p}}} m(\mathbf{p})\right] \ . \tag{3}$$

These two definitions are equivalent when the original diagnostic model is continuous (which is the case for models based on linear or logistic regression). When m is not continuous (3) gives a very good approximation of (2), and we therefore adopt (3) as the definition of $\hat{m} : \hat{P} \to \mathcal{I}_{[0,1]}$.

Example 2. *We will use the diagnostic model from the previous example, but now some patient data is missing:* $\mathbf{p}^A = (35, NA)$ *and* $\mathbf{p}^B = (NA, 1200)$*. According to (1) we define a new interval representation of patients*

$$\hat{\mathbf{p}}^A = ([35, 35], [0, 1500]) \ , \quad \hat{\mathbf{p}}^B = ([0, 100], [1200, 1200])$$

and compute diagnoses from uncertaintified models

$$\hat{m}_1(\hat{\mathbf{p}}^A) = \{m_1(p_1, p_2) : p_1 = 35, p_2 \in [0, 1500]\} = [0.088, 0.838]$$

and analogously $\hat{m}_1(\hat{\mathbf{p}}^B) = [0.6, 0.85]$*. It is easy to see that for the first patient it is hard to make a diagnosis, while for the second, despite the missing data, we can still say with high confidence that she has a malignant tumor.*

2.3. Aggregation of diagnoses

There are many different diagnostic models for ovarian tumor, and we wish to use this fact to improve the effectiveness of diagnosis. In our previous research we observed that different diagnostic models use different attributes describing the patient, and are therefore subject to different levels of uncertainty [10]. The main idea is thus to improve the final diagnosis by taking advantage of the models' diversity. Given n models $m_1, m_2, \dots m_n$ we construct a function Agg whose result is a new diagnosis that gathers and integrates information from the input models. Thanks to this interpretation we immediately see the relationship with the problem of group decision-making and information aggregation [12]. An n-argument aggregation operator is a mapping $\mathrm{Agg} : [0,1]^n \to [0,1]$ with the following properties [13]:

1. if $y_i \leq x_i$ for all $i \in 1, ..., n$, then $\mathrm{Agg}(y_1, ..., y_n) \leq \mathrm{Agg}(x_1, ..., x_n)$,
2. $\mathrm{Agg}(1, ..., 1) = 1\,,$
3. $\mathrm{Agg}(0, ..., 0) = 0\,.$

There are four main classes of aggregation operators: averaging, conjunctive, disjunctive and mixed. A detailed list of aggregation operators used in this research is given in Appendix A.1.

We use the interval representation introduced in the previous section. Thus, instead of numbers we will aggregate intervals. There are two possible modes of such aggregation. The first, called numerical, uses a single value that represents the whole interval (the most common choices are the interval's center, lower bound and upper bound). The interval mode, on the other hand, utilizes the whole of the interval information. Recent research has led to the construction of many aggregation methods, of both numerical and interval type [14, 15, 13, 16]. The most commonly used aggregation methods in group decision-making are based on the weighted arithmetic mean [12].

Example 3. *Continuing the previous examples, assume that there is a new blood marker ($D_3 = [0, 100]$) and it is used in a new diagnostic model*

$$m_2(\mathbf{p}) = 0.0025p_1 + 0.0075p_3\,.$$

New marker results were assessed for both patients with the following results: $\mathbf{p}^A = (35, NA, 5)$ *and* $\mathbf{p}^B = (NA, 1200, 90)$. *The new diagnostic model (after uncertaintification) yields* $\hat{m}_2(\hat{\mathbf{p}}^A) = [0.125, 0.125]$ *and* $\hat{m}_2(\hat{\mathbf{p}}^B) = [0.675, 0.925]$.

Having two different pieces of information, we can try to merge them into a single one which will be more reliable. What we know about the first patient is that the diagnostic models yielded $[0.088, 0.838]$ *and* $[0.125, 0.125]$ *as a suggested diagnosis. In this example we will present two modes of aggregation using a very simple and intuitive aggregation method, namely the arithmetic mean. In the mode based on numerical evaluation we choose the interval center as a representative. Calculation gives the following result:*

$$Agg\left(\hat{m}_1(\hat{\mathbf{p}}^A), \hat{m}_2(\hat{\mathbf{p}}^A)\right) = \frac{1}{2}(\frac{0.088 + 0.838}{2} + \frac{0.125 + 0.125}{2}) = 0.294\,.$$

Analogously for the second patient

$$Agg\left(\hat{m}_1(\hat{\mathbf{p}}^B), \hat{m}_2(\hat{\mathbf{p}}^B)\right) = \frac{1}{2}(\frac{0.6 + 0.85}{2} + \frac{0.675 + 0.925}{2}) = 0.763\,.$$

In the mode based on interval evaluation we use interval arithmetic. Calculation gives the following results:

$$\hat{Agg}\left(\hat{m}_1(\hat{\mathbf{p}}^A), \hat{m}_2(\hat{\mathbf{p}}^A)\right) = \left[\frac{0.088 + 0.125}{2}, \frac{0.838 + 0.125}{2}\right] = [0.107, 0.482]$$

and

$$\hat{Agg}\left(\hat{m}_1(\hat{\mathbf{p}}^B), \hat{m}_2(\hat{\mathbf{p}}^B)\right) = \left[\frac{0.6 + 0.675}{2}, \frac{0.85 + 0.925}{2}\right] = [0.638, 0.888]\,.$$

Thanks to the use of aggregation we have obtained new diagnoses which are less uncertain and make it easier to take a final decision.

One would expect that the medical diagnosis problem itself imposes some restrictions on the properties of the aggregation method. For example, if we make some assumptions about the quality of the original diagnostic models [17]:

1. All models are reliable – this leads to conjunctive aggregation methods,
2. At least one of the models is reliable – this leads to disjunctive aggregation methods,
3. The models provide independent information – this lead to counting based methods.

In practice, none of these holds, and the reality is somewhere in between. This may lead to the conclusion that averaging aggregation operators (which lie between the conjunctive and disjunctive operator families) can be expected to be a good solution to this problem. It is also noted that it is preferable to diagnose doubtful cases as malignant; this follows directly from the nature of the medical problem under consideration. A detailed theoretical discussion on this topic is not the subject of this paper.

2.4. Thresholding

In the medical decision-making problem, the final diagnosis must indicate whether a tumor is malignant (M) or benign (B). However, supporting a decision in a case where there is not enough information may lead to a wrong diagnosis. Therefore, we accept a situation in which no diagnosis recommendation is made (NA), but only in the last stage of the decision-making process. Since there are two different aggregation modes, we will need two different classes of thresholds:

1. numeric threshold $\tau : [0, 1] \to \{\mathrm{B}, \mathrm{M}, \mathrm{NA}\}$,
2. interval threshold $\hat{\tau} : \mathcal{I}_{[0,1]} \to \{\mathrm{B}, \mathrm{M}, \mathrm{NA}\}$.

For both classes a variety of methods can be constructed (see Appendix A.2).

Example 4. *Let us consider some diagnoses from the previous examples. For numerical modes we will use the simplest threshold method:*

$$\tau(x) = \begin{cases} M & \textit{if } x \geq 0.5 \\ B & \textit{if } x < 0.5 \end{cases} .$$

Application of this threshold leads to the following diagnosis recommendations: $\tau(0.294) = B$ *and* $\tau(0.763) = M$.
For interval input we check whether the whole interval is above or below the 0.5 *threshold:*

$$\hat{\tau}([a, b]) = \begin{cases} M & \textit{if } a \geq 0.5 \\ B & \textit{if } b \leq 0.5 \textit{ and } a < 0.5 \\ NA & \textit{otherwise} \end{cases} .$$

The following diagnoses are then made: $\hat{\tau}([0.107, 0.482]) = B$, $\hat{\tau}([0.638, 0.888]) = M$ *and* $\hat{\tau}([0.088, 0.838]) = NA$.

2.5. Summary of proposed approach

To sum up, our method assumes the use of effective and widely accepted diagnostic models. By the use of interval representation and the uncertaintification process, we have enabled these models to operate on incomplete data. The next step is to take advantage of the wide variety of diagnostic models for ovarian tumor. Thanks to the aggregation of decisions we were able to improve overall effectiveness and minimize the impact of incomplete data on

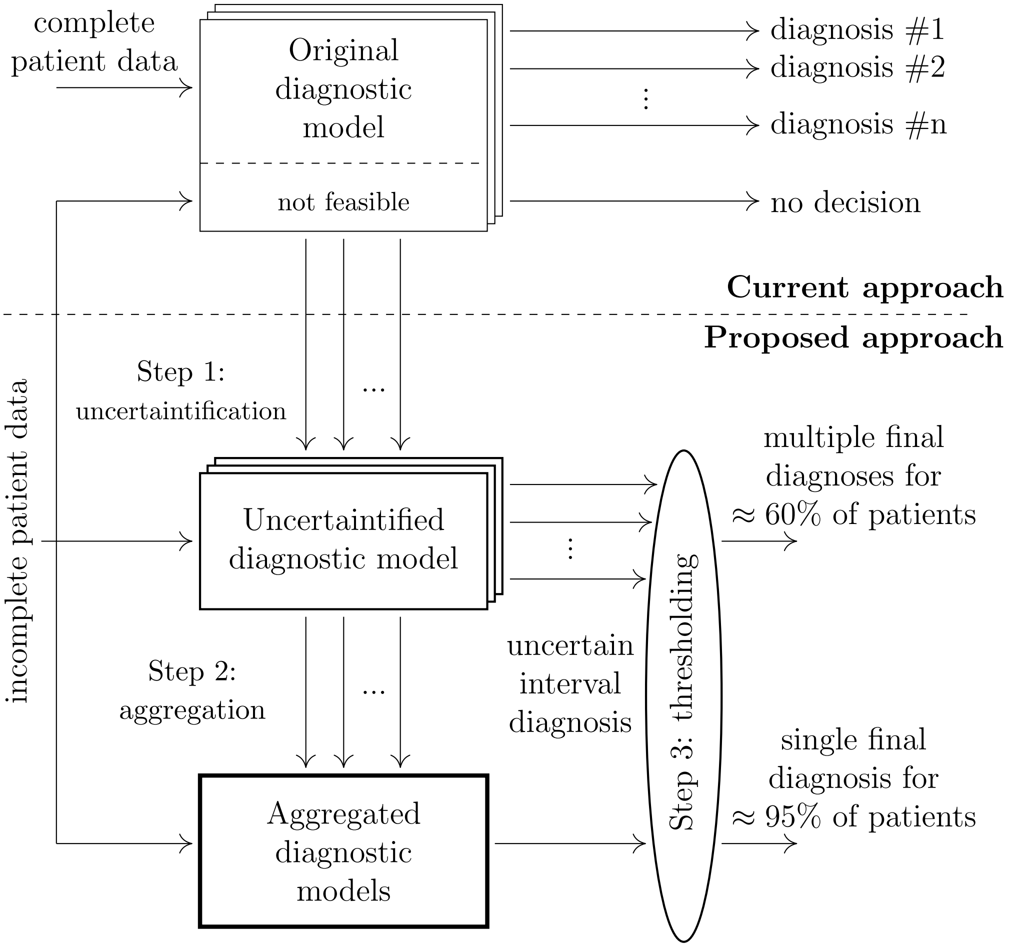


Figure 1: Graphical summary of current and proposed approaches. Rectangles represent diagnostic models at different stages. Vertical arrows represent diagnostic model transformations (uncertaintification and aggregation). The third step (thresholding) is depicted as an ellipse. Horizontal arrows represent the flow of patient data and diagnosis.

the final diagnosis. In the final step the interval diagnosis is converted to a form that can be understood by the decision-maker (the physician). The novelty is that we accept a situation in which no diagnosis recommendation is made, which protects us from making an unjustified wrong decision in cases where there is not enough information available. All three steps of the proposed approach are depicted in Figure 1. The next section concerns the implementation and evaluation of this approach in a real-life medical decision problem.

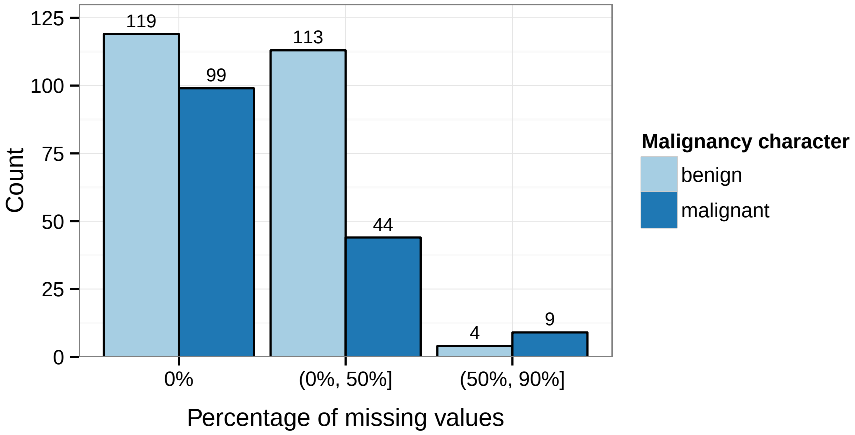


Figure 2: Distribution of patients in terms of percentage of missing values in the dataset.

3. Evaluation

The following section describes a dataset as well as an evaluation procedure. Furthermore, criteria for assessing the quality of classification performance are discussed. The last subsection contains a technical description of the research procedure.

3.1. Subject of the evaluation

The study group consists of 388 patients diagnosed and treated for ovarian tumor in the Division of Gynecological Surgery, Poznan University of Medical Sciences, between 2005 and 2015. Among them, 61% were diagnosed with a benign tumor and 39% with a malignant one. Moreover, 56% of the patients had no missing values in the attributes required by diagnostic scales, 40% had a percentage of missing values in the range $(0\%, 50\%]$, and the remainder had more than 50% missing values. The distribution of the percentages of missing values is shown in Figure 2. The subset of the dataset is described in detail in [6]. More information about the data format can be found in the technical supplement (see section 3.5).

For the evaluation process we selected six diagnostic models: two scoring systems [18, 19] and four regression models [20, 21, 22]. Table 1 presents the models and the patient features used by them. The attributes are divided into two groups – the first is always available, and the second may have some missing values. The models are subjected to the uncertaintification process as described in section 2.1.

	diagnostic models					
attribute	SM	Alc.	LR1	LR2	Tim.	RMI
	m_1[18]	m_2[19]	m_3[20]	m_4[20]	m_5[21]	m_6[22]
age	-	-	✓	✓	-	✓
menopausal status	✓	-	-	-	✓	✓
pain during examination	-	-	✓	-	-	-
hormonal therapy	-	-	✓	-	-	-
hysterectomy	-	-	-	-	-	✓
ovarian cancer in family	-	-	✓	-	-	-
lesion volume	✓	-	✓	-	-	-
internal cyst walls	✓	-	✓	✓	-	-
septum thickness	✓	-	-	-	-	-
echogenicity	✓	✓	-	-	-	-
localisation	✓	-	-	-	-	✓
ascites	✓	-	✓	✓	-	✓
papillary projections	-	✓	-	-	✓	-
solid element size	-	✓	✓	✓	-	✓
blood flow location	-	✓	✓	✓	-	-
resistance index	-	✓	-	-	-	-
acoustic shadow	-	-	✓	✓	-	-
amount of blood flow	-	-	✓	-	✓	-
CA-125 blood marker	-	-	-	-	✓	✓
lesion quality class	-	-	-	-	-	✓

Table 1: Attributes used in the most common preoperative diagnostic models. Features in the second group may have missing values.

For the aggregation step we considered four groups of aggregation operators: weighted averages (including r-means), OWA, Chocquet and Sugeno integrals, and t-operations (t-norms, t-conorms and generalized conjunctions/disjunctions). Within each group, we investigated two subgroups of aggregation operators, differing in whether they aggregate whole intervals or the numerical representatives of intervals. The aggregation operators are described in detail in Appendix A.1.

Finally, we examined several thresholding strategies. For the resulting interval or numerical values we checked how they differ from the value 0.5; that is, whether they are greater or lower than $0.5 \pm \epsilon$, where $\epsilon \geq 0$. For the resulting intervals, we also tested the largest common part strategy. That is, interval intersections were calculated with three intervals indicating regions associated with benign, malignant and unknown (NA) output: $[0, 0.5 - \epsilon]$, $[0.5 - \epsilon, 0.5 + \epsilon]$ and $[0.5 + \epsilon, 1]$. We examined which intervals have the largest common part, as well as whether one intersected region is larger than the sum of the other two. The thresholding strategies are described in detail in Appendix A.2.

3.2. Assumptions on dataset partition

The evaluation procedure was based on the classic division of data into training and test sets. The initial dataset has very few patients with missing attributes for some levels of missing data. If the data were divided approximately evenly, this would lead to a situation in which at the stage of training and/or testing there would be discontinuities in the distribution of data missingness – while our goal is to develop a working algorithm for each level of missing data. An alternative is to enlarge the dataset by including new patients; however, such a solution is very time-consuming and costly. Therefore we chose a different solution, as described below.

The test set consists of patients with real missing data and some proportion of patients with a complete set of features. The training set, on the other hand, is constructed on a set of patients with a complete set of attributes, and the incompleteness is then simulated. In the simulations we assumed that the data are missing at random – this is because it is impossible to reflect the true process by which data come to be missing during the examinations. The actual distribution of data missingness in the patient attributes is unknown, so we decided that in the training phase different levels of data missingness would be simulated uniformly. Consequently, both training and test sets have no discontinuities in the distribution of data missingness.

In addition, the real distribution of tumor malignancy in the population is also unknown. Some statistics on this issue can be found in [5], where the authors list almost all diagnostic models for ovarian tumor classification with the distribution of malignancy in particular study groups. The benign/malignant ratio varies widely among the groups, and there is no guarantee that patients are not duplicated among different groups. Therefore, in the simulation process, during repeated random sampling of patients and obscuring of data, we assumed that the distribution of tumor malignancy is equal, so we randomly selected the same number of patients with benign and malignant tumors.

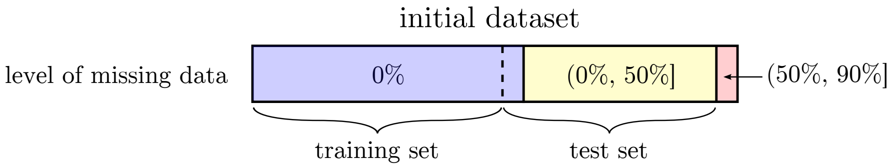


Figure 3: The division of the dataset. Patients with more than 50% missing values were not included in the experiment.

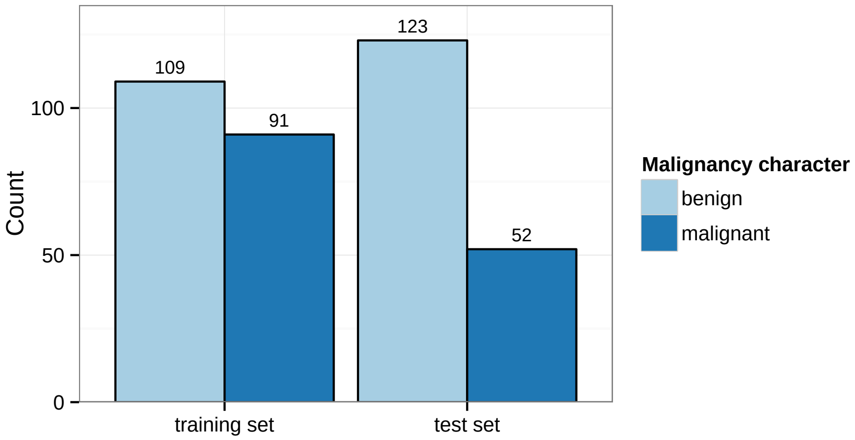


Figure 4: Class distribution in the training and test sets.

3.3. Evaluation procedure

With regard to the assumptions mentioned in section 3.2, the input dataset described in 3.1 was divided as follows. In the training set there are 200 patients with no missing values in the attributes required by the diagnostic models. The test set consisted of the remaining 18 patients with no missing values and those who had missing values in the range $(0\%, 50\%]$. As a result, the test set consisted of 175 patients. The aforementioned subgroups of 200 and 18 patients had the same proportion of benign to malignant cases. Patients with more than 50% missing values were excluded from the study. The dataset partition is presented visually in Figure 3 and Figure 4.

The aim of the training phase is to optimize the parameters of the aggregation operators and thresholding strategies on different simulated percentages of missing features. The levels of missing data were set to vary from 0% to 50% with a step size of 5%. For each level, 1000 repetitions were made of the following procedure. Firstly, 75 patients with benign tumors and 75 patients with malignant tumors were randomly selected from the training set. Secondly, a given percentage of patients' features from the second group of attributes were obscured (removed). Next, with such input, the uncertaintified diagnostic models calculated interval-valued diagnoses. Finally, the aggregation operators and thresholding strategies calculated final diagnoses. Performance measures were calculated according to the assumptions in Section 3.4. All results are averaged over the repetitions and the levels of missing data. To avoid overfitting, the aggregation operators and thresholding strategies are optimized on a reasonable set of numerical parameters, selected by an expert. Each step of the training phase is presented in Fig. 5. The result of the training phase is a set of optimized aggregation operators with thresholding strategies which have good performance on simulated missing data.

In the testing phase, the optimized aggregation operators and thresholding strategies are examined on the test set. This step checks the performance of these aggregation operators on data with the actual missing values. Again, the performance is calculated according to Section 3.4. To estimate uncertainty of the performance values obtained on the test set, we performed stratified bootstraping with 500 replications.

3.4. Criteria of performance evaluation

The aim of the evaluation is to choose such an aggregation operator and thresholding strategy which provide an accurate diagnosis of malignant cases with the highest possible decisiveness. The performance of learning algorithms can be expressed by many state-of-the-art measures, such as accuracy, sensitivity and specificity. In the problem considered here, the desired solution should ensure very high sensitivity and high specificity. Moreover, in a few cases the results of diagnostic models are ambiguous, so the classification method should not perform a classification by chance – in such a case the patient should be referred to an experienced gynecologist. Hence we accept a situation where a few percent of the patients still have no diagnosis recommendation (decisiveness lower than 100%). Since the selection of an appropriate unified performance measure is a difficult task (see [23]), the cost matrix approach was considered.

Table 2 presents the costs of possible decisions made by a classifier. The correct classification of tumors, as *true positives* and *true negatives*, comes with zero cost. The highest cost is associated with *false negatives*, when a patient has a malignant tumor and the prediction indicates that it is benign. The cost of a *false positive* is set to be two times smaller than that of a *false*

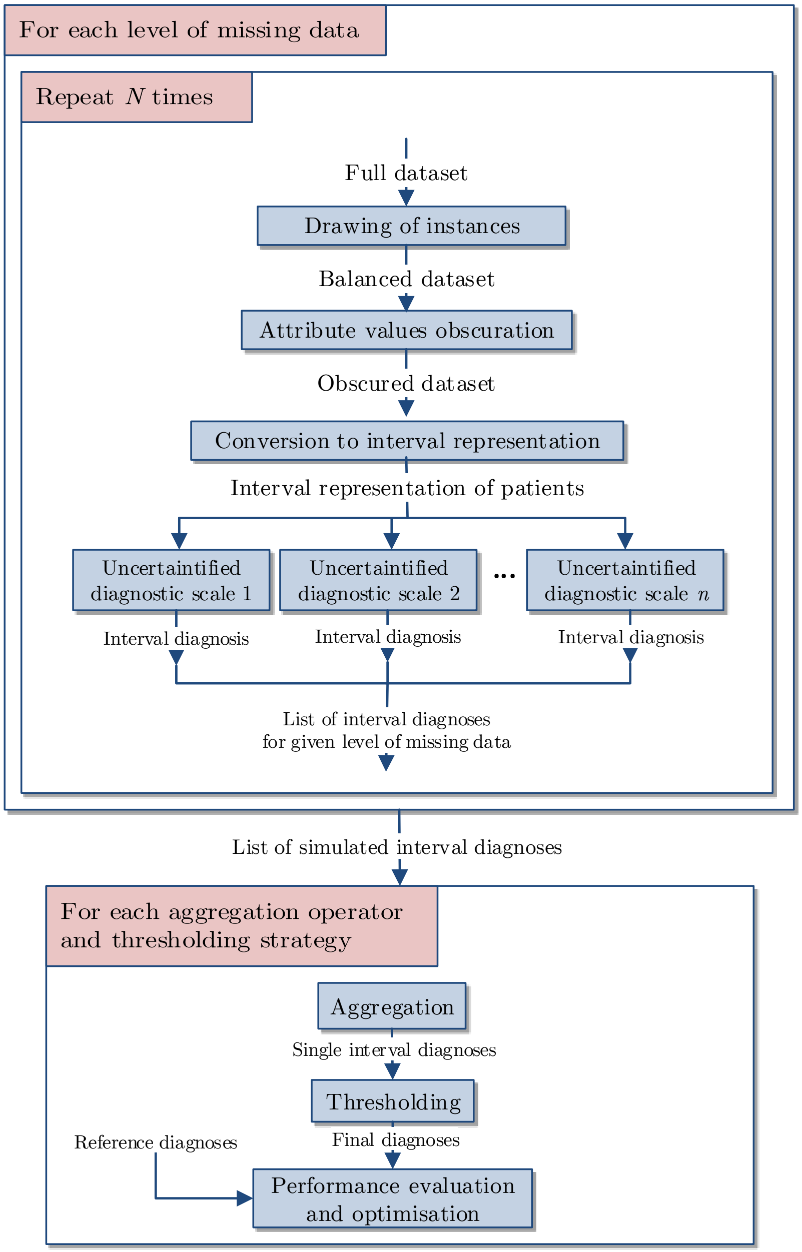

For each level of missing data
Repeat N times
Full dataset
Drawing of instances
Balanced dataset
Attribute values obscuration
Obscured dataset
Conversion to interval representation
Interval representation of patients
Uncertaintified diagnostic scale 1
Uncertaintified diagnostic scale 2
...
Uncertaintified diagnostic scale n
Interval diagnosis
Interval diagnosis
Interval diagnosis
List of interval diagnoses for given level of missing data
List of simulated interval diagnoses
For each aggregation operator and thresholding strategy
Aggregation
Single interval diagnoses
Thresholding
Reference diagnoses
Final diagnoses
Performance evaluation and optimisation

Figure 5: Visualization of the training phase. Data flow is represented as arrows, and boxes represent operations on data.

		predicted		
		benign	malignant	NA
actual	benign	0	2.5	1
	malignant	5	0	2

Table 2: Cost matrix. The costs were assigned based on expert gynecologists' opinions.

negative, since unnecessary surgery is still dangerous for a patient, but there is a much greater chance of recovery. There is also a certain difference in costs when a classifier does not know which class should be assigned. The costs of no decision (NA) are lower than *false positive*, since the patient is referred to an experienced gynecologist who is still able to make a good decision. However a further misclassification is not ruled out, so in case of no prediction, the cost when the tumor is malignant is two times greater than the cost when it is benign.

3.5. Technical issues

The statistical evaluation, as well as the implementation of the proposed methodology, were performed using R software, version 3.1.2 [24]. All scripts, documentation and non-sensitive data are available on the GitHub: `http://ovaexpert.github.io/ovarian-tumor-aggregation`. Due to the extensive amount of computation required, all calculations were performed with the use of the Microsoft Azure cloud service.

4. Results

This section presents and discusses in detail the results obtained during the training and testing phases.

In the training phase eight groups of aggregation operators and four thresholding strategies were considered and optimized to minimize the total cost obtained according to the cost matrix. As a result we obtained a set of aggregators with optimized parameters, and the best among them were chosen. The top tree aggregation operators and thresholding strategies within each group are listed in Table 3.

Figure 6 summarizes an experiment conducted on a training data set that was subjected to the process of obscuration. First, the original diagnostic

		Performance measures with 95% CI			
No.	Operator parameters	Total cost	Decisiveness	Sensitivity	Specificity
Integrals in interval mode given by (A.15) and (A.16)					
1	Choquet, μ_{AUC}, $\hat{\tau}_{\text{cen},0.025}$	80.0 (±28.8)	92.0 (±4.0)	82.6 (±11.0)	93.0 (±4.6)
2	Choquet, μ_{card}, $\hat{\tau}_{\text{cen},0.025}$	80.0 (±27.8)	92.0 (±4.1)	84.8 (±10.3)	91.3 (±5.3)
3	Sugeno, μ_{card}, $\hat{\tau}_{\text{cen},0.025}$	80.0 (±26.1)	87.4 (±4.9)	90.9 (±8.0)	89.0 (±6.2)
Integrals in numerical mode given by (A.8) and (A.9)					
4	Choquet, rep_{cen}, μ_{AUC}, $\tau_{0.025}$	80.0 (±28.8)	92.0 (±4.0)	82.6 (±11.0)	93.0 (±4.6)
5	Choquet, rep_{cen}, μ_{card}, $\tau_{0.025}$	80.0 (±27.8)	92.0 (±4.1)	84.8 (±10.3)	91.3 (±5.3)
6	Sugeno, rep_{min}, μ_{card}, $\tau_{0.0}$	87.5 (±31.8)	100.0 (–)	86.5 (±8.6)	82.9 (±6.9)
Weighted means in interval mode given by (A.13)					
7	ω_{wid}, $r = 2$, $\hat{\tau}_{\text{cen},0.025}$	75.5 (±26.2)	97.1 (±2.6)	91.8 (±7.2)	84.3 (±6.2)
8	ω_{em}, $r = 3$, $\hat{\tau}_{\text{cen},0.0}$	77.5 (±28.1)	100.0 (–)	88.5 (±8.7)	84.6 (±6.3)
9	ω_1, $r = 2$, $\hat{\tau}_{\text{cen},0.025}$	79.0 (±27.5)	94.3 (±3.2)	91.7 (±7.3)	84.6 (±6.4)
Weighted means in numerical mode given by (A.6)					
10	rep_{min}, ω_{ep}, $r = 3$, $\tau_{0.0}$	72.0 (±27.5)	97.1 (±2.6)	90.0 (±8.6)	86.7 (±5.9)
11	rep_{cen}, ω_{ep}, $r = 3$, $\tau_{0.0}$	74.5 (±27.9)	97.1 (±2.6)	90.0 (±8.6)	85.8 (±5.9)
12	rep_{min}, ω_{wid}, $r = 3$, $\tau_{0.025}$	78.0 (±30.0)	94.3 (±3.4)	85.7 (±9.3)	89.7 (±5.5)
Ordered Weighted Average (OWA) operators in interval mode given by (A.14)					
13	ω_{dec}, π_{cen}, $\hat{\tau}_{\text{cen},0.025}$	70.0 (±29.1)	94.9 (±3.4)	90.2 (±8.3)	87.8 (±5.9)
14	ω_{dec}, π_{min}, $\hat{\tau}_{\text{cen},0.025}$	72.0 (±29.3)	96.6 (±2.8)	90.2 (±8.3)	86.4 (±5.9)
15	ω_{dec}, π_{wm}, $\hat{\tau}_{\text{cen},0.025}$	73.5 (±28.4)	94.9 (±3.1)	90.0 (±8.5)	87.1 (±6.2)
Ordered Weighted Average (OWA) operators in numerical mode given by (A.7)					
16	rep_{cen}, ω_{dec}, π_{cen}, $\tau_{0.025}$	70.0 (±29.1)	94.9 (±3.4)	90.2 (±8.3)	87.8 (±5.9)
17	rep_{cen}, ω_{dec}, π_{min}, $\tau_{0.025}$	72.0 (±29.3)	96.6 (±2.8)	90.2 (±8.3)	86.4 (±5.9)
18	rep_{cen}, ω_{dec}, π_{wm}, $\tau_{0.025}$	73.5 (±28.4)	94.9 (±3.1)	90.0 (±8.5)	87.1 (±6.2)
t-operation based operators in interval mode given by (A.17)					
19	s_{max}, $\alpha = 0.25$, $\hat{\tau}_{\text{cen},0.025}$	78.0 (±26.6)	94.3 (±3.4)	91.8 (±7.0)	84.5 (±6.6)
20	t_{min}, $\alpha = 0.25$, $\hat{\tau}_{\text{max},0.025}$	89.5 (±28.5)	94.9 (±3.1)	89.8 (±8.8)	82.1 (±6.9)
21	t_{min}, $\alpha = 1.0$, $\hat{\tau}_{\text{max},0.0}$	100.0 (±35.0)	100.0 (–)	73.1 (±12.5)	90.2 (±5.2)
t-operation based operators in numerical mode given by (A.12)					
22	rep_{cen}, s_{max}, $\alpha = 0.25$, $\tau_{0.025}$	82.0 (±27.5)	94.9 (±2.8)	89.8 (±8.4)	84.6 (±6.4)
23	rep_{max}, t_{min}, $\alpha = 0.25$, $\tau_{0.025}$	89.5 (±28.5)	94.9 (±3.1)	89.8 (±8.8)	82.1 (±6.9)
24	rep_{min}, s_{prod}, $\alpha = 0.25$, $\tau_{0.025}$	95.0 (±29.7)	93.7 (±3.2)	87.5 (±8.9)	82.8 (±7.2)

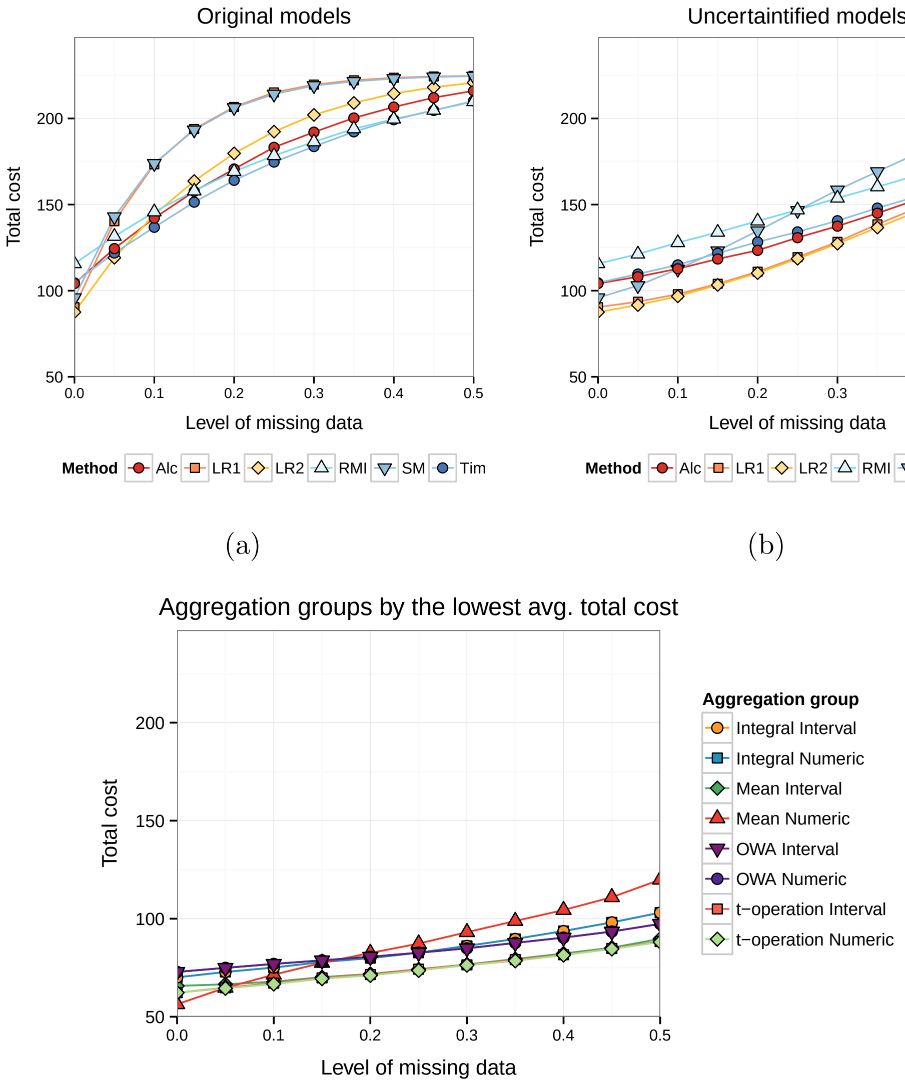
Table 3: Performance measures for the top three aggregation operators and thresholding strategies within each group. Abbreviations: Dec. – Decisiveness, Sen. – Sensitivity, Spec. – Specificity, Acc. – Accuracy. All measures, along with bootstrap percentile 95% confidence intervals, are achieved in the test set. The decisiveness, sensitivity and specificity are in percentage values.

models were run[1] on that data and the total cost was calculated according to the cost matrix, for each model and each level of missing data. The graph of total costs is shown in Fig. 6(a). As expected, the cost grows rapidly with increasing level of missing data. This is because classical models are not able to make a diagnosis when some attributes' values are not available, thus they fail to predict any value (NA). Next, the diagnostic models were uncertaintified and again run on the training data with missing values. The results can be seen in Fig. 6(b). This time the total cost for each of the models grows more slowly. This demonstrates that uncertaintification itself makes it possible to reduce the impact of data incompleteness on the effectiveness of diagnosis. Finally, in Fig. 6(c) we can observe the cost of diagnosis obtained via aggregation. The diagram shows one arbitrarily chosen aggregation operator for each group. The total cost is much lower than in cases (a) and (b) for each level of missing data, and its growth is small. Consequently, the training phase allows us to obtain a set of aggregation operators that are significantly better for ovarian tumor diagnosis than single diagnostic models, for each level of missing data.

Next, the results obtained during the training phase were verified on the test data set with actual (not simulated) missing values. Figure 7 presents the total cost for (a) original models, (b) uncertaintified models, and (c) aggregation operators. Again, the cost is highest for original models and lowest for aggregation operators. These results confirm the claim that aggregation operators are a good tool for diagnosis in the presence of missing data.

A detailed analysis of diagnostic models and the best aggregation operators is depicted in Fig. 8. The accuracy, sensitivity, specificity and decisiveness of each of the methods are presented. It can be seen that for the original models accuracy, sensitivity and specificity are quite high, but decisiveness is low – in many cases a patient is not diagnosed at all. On the other hand, aggregation operators produce equal or even higher values of accuracy, sensitivity and specificity, while at the same time achieving high decisiveness – a diagnosis is unavailable for fewer than 10% of patients. These are very good results, showing that aggregation is a promising way of improving the quality of medical diagnosis.

[1] An original model may still classify if missingness of patients' features does not concern attributes used by the model. Although the original and uncertaintified models are not the subject of the optimisation in the proposed approach, they are plotted for comparison with the aggregation.

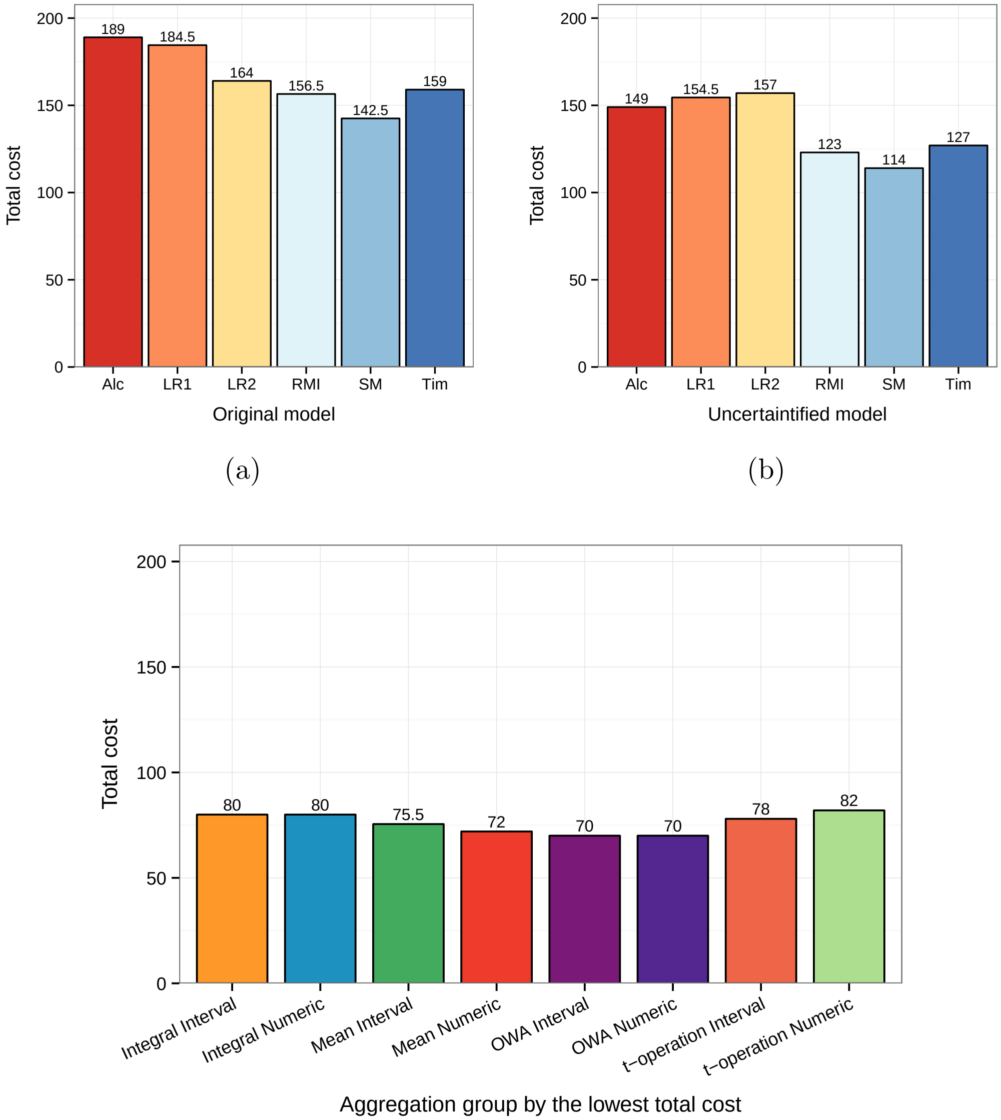


(a)

(b)

(c)

Figure 6: Simulation results from (a) original models, (b) uncertaintified models and (c) aggregation groups. The aggregation groups show strategies with the lowest total cost achieved on the training set.

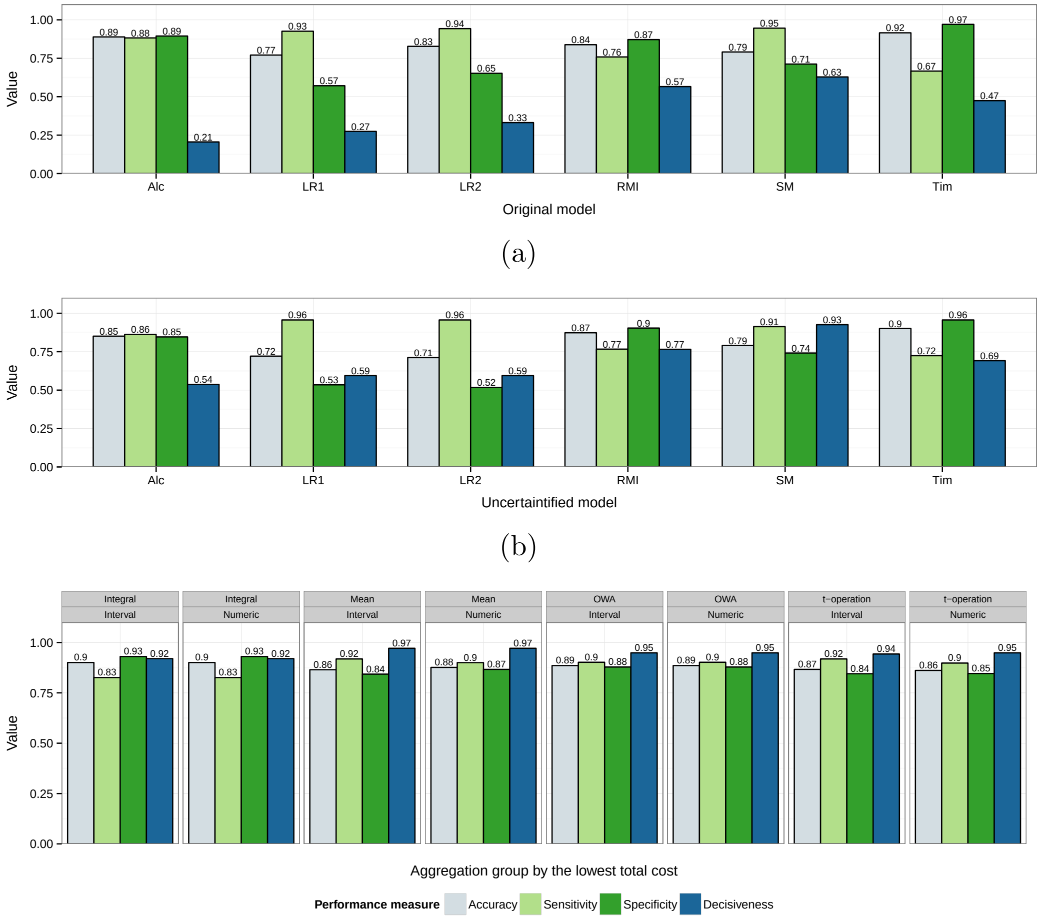


(a)

(b)

(c)

Figure 7: Comparison of total cost performance on the test set among (a) original models, (b) uncertaintified models, and (c) aggregation groups.



(a)

(b)

(c)

Figure 8: Comparison of accuracy, sensitivity, specificity and decisiveness on the test set among (a) original models, (b) uncertaintified models, and (c) aggregation groups.

5. Discussion

There were at least two reasons for carrying out the present study. The first was to explore a new, general approach to dealing with missing data through uncertaintification and aggregation of existing methods. This goal was successfully achieved, and the results of the experiments confirmed that aggregation operators are a good tool to support decision-making in the presence of incomplete information.

The second objective was to incorporate aggregation methods into the

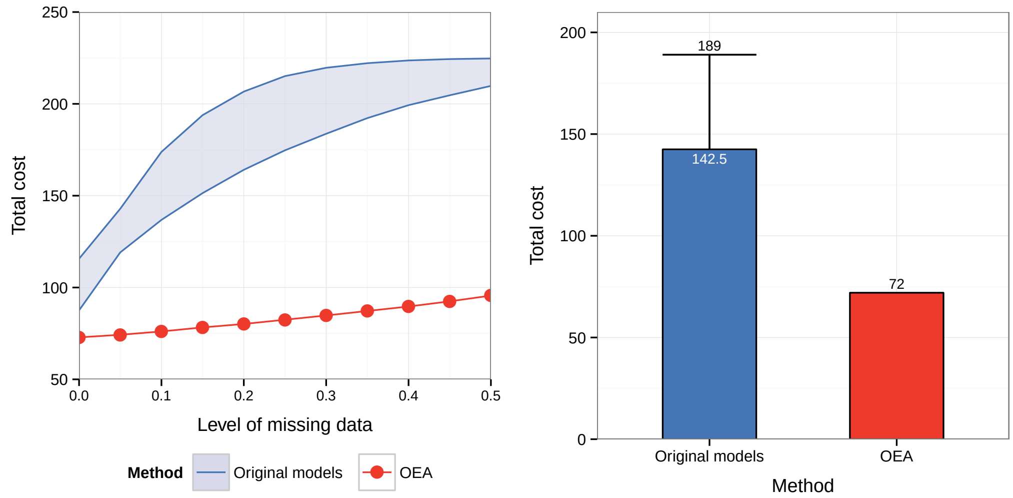
OvaExpert system, to allow ovarian tumor diagnosis when some of the data are missing. OvaExpert is an innovative system based on machine learning techniques and computational intelligence. The system is designed to integrate present knowledge about ovarian tumors (models, scoring systems, reasoning schemes, etc.) into a single computer-aided system. Its main aim is to equip a physician with a convenient tool to collect and manage patient data, to minimize the negative impact of incomplete data on the final diagnosis, to improve the reliability and efficacy of the diagnosis, and finally to present the result in a way that gives maximum information to the doctor.

One of the diagnostic modules in OvaExpert is based on aggregation of diagnostic models. To this end we need to choose only one aggregation operator that is best suited to our problem. To select the best aggregation operator from those returned in the training and testing phases, we require that the following conditions be satisfied:

- sensitivity $\geq$ 90%,
- specificity $\geq$ 80%,
- sensitivity $>$ specificity,
- decisiveness $<$ 100%.

The first two rules filter out aggregation operators with high sensitivity and specificity values. The third rule reflects the fact that in a medical context sensitivity is more important than specificity. Since these two measures are correlated there may be some models (aggregation operators) that trade off sensitivity for specificity – we reject such models. Finally, we exclude models with 100% decisiveness, since we do not wish to impose diagnoses that lack sufficient justification. In this case no decision, leading to further examinations, is better than a wrong decision.

The operator that was chosen for OvaExpert will be further referred to as OEA (OvaExpert Aggregator). It is an OWA operator defined by (A.7) with the weight vector ω_{dec}, $\mathrm{rep}_{\mathrm{cen}}$ as representative selector, $\tau_{0.025}$ as threshold and $\pi_{\min}$ is used to order input values (see Appendix A for detailed definitions). The total cost of this operator in comparison with the original diagnostic models is depicted in Fig. 9a (on the training dataset) and Fig. 9b (on the test dataset). OEA is significantly better than all of the other diagnostic models. This was verified with McNemar's statistical test, and the results are given in Table 4.



(a) (b)

Figure 9: A comparison of total costs between the original diagnostic models and the selected aggregation strategy in the training (a) and test (b) phases. The shaded area in (a) indicates lower and upper bounds of the total cost of the original models. The vertical range on the first bar in (b) indicates lower and upper bounds of the total cost of all original models.

6. Conclusions

The main contribution of the paper is the solution provided to the problem of making decisions under incomplete information. Our approach is based on interval modeling, uncertaintification and aggregation of existing models. We have not only presented the theoretical concept, but also conducted exhaustive testing and provided a framework written in R that can be used by other researchers in many disciplines, not only in medicine. Apart from the general method of dealing with missing data, we have developed OvaExpert, a complex system for diagnosis support. The results presented here form part of that system.

The basic conclusion of our study is that with our approach we can obtain better results in ovarian tumor diagnosis than when using known diagnostic models. This is especially evident when diagnosis is based on incomplete data. Using aggregation we can achieve a synergy effect and become able to cope with quite a large amount of missing data (up to 50%). Selected method (OEA) is able to give proper diagnosis despite missing data. Total cost of 72

		Original models					
		Alc.	LR1	LR2	RMI	SM	Tim.
Uncertaintified models	Alc.	< 0.001	< 0.001	< 0.001	0.834	0.472	0.723
	LR1	< 0.001	< 0.001	< 0.001	0.406	0.080	1.000
	LR2	< 0.001	< 0.001	< 0.001	0.366	0.060	0.935
	RMI	< 0.001	< 0.001	< 0.001	< 0.001	0.001	< 0.001
	SM	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
	Tim.	< 0.001	< 0.001	< 0.001	0.001	0.017	< 0.001
	OEA	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

Table 4: McNemar's test with Benjamini-Hochberg correction between the original diagnostic models and the uncertaintified models with the selected aggregation strategy. It can be observed that the uncertaintified models significantly outperform the corresponding original models. Moreover, OEA significantly outperforms the original models ($\alpha = 0.05$).

is very low compared to original diagnostic models (142.5-189). Moreover, very high sensitivity and specificity proves that proposed approach can be applied in real medical practice.

We are aware that such evaluation would benefit from the usage of other external datasets. One of the reasons for development of the OvaExpert system is to enable us to gather new medical data. We established a cooperation with some medical centres from Europe. Our goal is to create new, diversified datasets which are now not publicly available in this research field.

Our study was limited to the case of supporting ovarian tumor diagnosis, but it can be adapted to other decision-making problems for which models equivalent to diagnostic models are available. Our further research is aimed towards improving the aggregation methods so that they better suit the specifics of the patient's description. Moreover, considering the high cost of medical examinations, we wish to be able to distinguish attributes that must be filled from those that can remain empty without a significant loss of decisiveness. In addition, we have begun research into the possibility of using the developed methods in cardiology and in economics.

Appendix A. Evaluated approaches

This section describes all groups of aggregation operators and thresholding strategies with all required parameters (such as the method of weight calculation or selection of interval representatives) used in the evaluation.

Appendix A.1. Aggregation operators

This subsection lists all aggregation methods evaluated in our research. There are four groups of operators: r-means, OWA, integrals and t-operations. Each group is represented both in numerical and interval aggregation mode.

Appendix A.1.1. Weight calculation strategies

Many aggregation operators involve assigning appropriate weights to input values. The problem is the same regardless of the mode of aggregation. Thus, we combine the description of different weight calculation strategies into one subsection.

The following weight calculation strategies were implemented in this research:

- constant value:
$$\omega_1([a,b]) = 1\,, \tag{A.1}$$
- interval length:
$$\omega_{\text{wid}}([a,b]) = 1 - (b-a)\,, \tag{A.2}$$
- interval endpoint distance from 0.5:
$$\omega_{\text{ep}}([a,b]) = \begin{cases} 0 & \text{if } a \leq 0.5 \leq b \\ 2(a-0.5) & \text{if } a \geq 0.5 \\ 2(0.5-b) & \text{otherwise}\,. \end{cases} \tag{A.3}$$
- interval center distance from 0.5:
$$\omega_{\text{em}}([a,b]) = 2 \cdot |0.5 - \frac{a+b}{2}|\,, \tag{A.4}$$
- lower and upper bounds of interval and interval center (ω_{min}, ω_{max} and ω_{cen}, respectively),
- combined interval center and width
$$\omega_{\text{wm}}([a,b]) = \frac{a+b}{2} \cdot (1-(b-a))\,. \tag{A.5}$$

Appendix A.1.2. Numerical mode

Aggregation methods that operate in this mode use a single value that represents the whole interval. Such a representative of the interval $\hat{x}$ is denoted by $\mathrm{rep}(\hat{x})$. We evaluated the three most obvious representatives, namely the lower ($\mathrm{rep}_{\mathrm{min}}$) and upper ($\mathrm{rep}_{\mathrm{max}}$) bound and center of the interval ($\mathrm{rep}_{\mathrm{cen}}$). This behavior simplifies the problem to classical non-interval aggregation. For more information about the presented aggregation methods we refer the reader to [12].

Weighted r-means. The weighted mean is probably the most commonly used method of aggregation. r-means generalize this concept by the using r-th power of each argument (for $r = 1$ the r-mean becomes the classical weighted mean). For weighted means, the selection of weights is crucial and determines the final outcome of the aggregation. The general formula for weighted r-mean is the following:

$$\mathrm{Agg}_{\mathrm{mean}}(\hat{x}_1, \hat{x}_2, ..., \hat{x}_n) = \sqrt[r]{\frac{\sum_{i=1}^{n} \omega(\hat{x}_i) \cdot \mathrm{rep}(\hat{x}_i)^r}{\sum_{i=1}^{n} \omega(\hat{x}_i)}}\,. \tag{A.6}$$

Ordered weighted average (OWA). This class of aggregation operators was developed by Yager in 1988 [25]

$$\mathrm{Agg}_{\mathrm{OWA}}(\hat{x}_1, \hat{x}_2, ..., \hat{x}_n) = \frac{\sum_{i=1}^{n} \omega_i \cdot \mathrm{rep}(\hat{x}_{\pi(i)})}{\sum_{i=1}^{n} \omega_i}\,. \tag{A.7}$$

In contrast to the arithmetic mean, in the *ordered weighted average* the weight vector is constant, while the input variables are ordered with respect to a certain criterion. Our implementation of OWA supports the ordering of input values with respect to any of the weights introduced in Appendix A.1.1. Such an ordering obtained from weight ω is denoted by π_ω. We used the following predefined weight vectors:

- $[0, 0.25, 0.5, 0.5, 0.75, 1]$ – denoted by ω_{inc},
- $[1, 0.75, 0.5, 0.5, 0.25, 0]$ – denoted by ω_{dec},
- $[0.1, 0.5, 1, 1, 0.5, 0.1]$ – denoted by ω_{hill},
- $[1, 0.5, 0.1, 0.1, 0.5, 1]$ – denoted by ω_{pit}.

Choquet and Sugeno integrals. These are two classes of aggregation operators defined with the use of the measure μ. Their main advantage is that they are able to model interactions between input variables.

The Choquet integral is given by

$$\mathrm{Agg}_{\mathrm{Cho}}(\hat{x}_1, \hat{x}_2, ..., \hat{x}_n) = \sum_{i=1}^{n} (\mu(H_i) - \mu(H_{i-1})) \cdot \mathrm{rep}(\hat{x}_{\pi(i)}) \tag{A.8}$$

and the Sugeno integral by

$$\mathrm{Agg}_{\mathrm{Sug}}(\hat{x}_1, \hat{x}_2, ..., \hat{x}_n) = \max_{i=1 \text{ to } n} \left[\min(\mu(H_i), \mathrm{rep}(\hat{x}_{\pi(i)})\right] \tag{A.9}$$

where $H_i = \{\pi(1), \pi(2), ..., \pi(i)\}$, π is a non-decreasing permutation of input variables and μ is a measure. The following measures are implemented in this research:

- set cardinality
$$\mu_{\mathrm{card}}(H) = \frac{|H|}{n} \tag{A.10}$$
- additive measure
$$\mu_{\mathrm{AUC}}(\{h_1, h_2, ...\}) = \sum_{i=1} \mu(\{h_i\}) \tag{A.11}$$
where the measure of a singleton was determined using the area under the ROC curve (AUC) of the original diagnostic models (the greater the AUC, the higher the measure).

Triangular operations. The last class of numerical aggregation operators is based on triangular operations, namely:

- t-norms (for $\alpha = 1$),
- t-conorms (for $\alpha = 1$),
- soft t-norms (for $\alpha < 1$),
- soft t-conorms (for $\alpha < 1$).

This class of operators is given by the formula

$$\mathrm{Agg}_{\Phi}(\hat{x}_1, ..., \hat{x}_n) = \frac{1-\alpha}{n} \sum_{i=1}^{n} \mathrm{rep}(\hat{x}_{\pi(i)}) + \alpha \cdot \Phi(\mathrm{rep}(\hat{x}_1), ..., \mathrm{rep}(\hat{x}_n)) \,. \tag{A.12}$$

Appendix A.1.3. Interval mode

Interval mode utilizes the whole of the interval information. The literature contains two approaches to adapting numerical aggregation strategies to operate on interval data. The first involves use of interval arithmetic, and the second the application of the original operator to the lower and upper bound separately. Both methods are presented.

Interval weighted r-means. These aggregation operators are obtained from numerical r-means by the use of interval arithmetic for all calculations. The formula is as follows

$$\hat{\mathrm{Agg}}_{\mathrm{mean}}(\hat{x}_1, \hat{x}_2, ..., \hat{x}_n) = \sqrt[r]{\frac{\sum_{i=1}^{n} \omega(\hat{x}_i) \times \hat{x}_i^r}{\sum_{i=1}^{n} \omega(\hat{x}_i)}} \tag{A.13}$$

but now $\sum$ denotes the sum of intervals, and multiplication (division) is replaced by multiplication (division) of interval by a constant.

Interval OWA. A generalization of OWA to operate on intervals was proposed by Yager [14, 26]

$$\hat{\mathrm{Agg}}_{\mathrm{OWA}}(\hat{x}_1, ..., \hat{x}_n) = [\mathrm{Agg}_{\mathrm{OWA}}(\underline{x}_1, ..., \underline{x}_n), \mathrm{Agg}_{\mathrm{OWA}}(\overline{x}_1, ..., \overline{x}_n)] \ . \tag{A.14}$$

The main idea is to apply an OWA operator for the lower and upper bounds of the input intervals separately, and to form an interval from the two results.

Interval Choquet and Sugeno integrals. An analogous approach was applied to define interval Choquet and Sugeno integrals [14]. They are defined by

$$\begin{aligned}\hat{\mathrm{Agg}}_{\mathrm{Cho}}(\hat{x}_1, \hat{x}_2, ..., \hat{x}_n) = [&\mathrm{Agg}_{\mathrm{Cho}}(\underline{x}_{\pi(1)}, \underline{x}_{\pi(2)}, ..., \underline{x}_{\pi(n)}), \\ &\mathrm{Agg}_{\mathrm{Cho}}(\overline{x}_{\pi(1)}, \overline{x}_{\pi(2)}, ..., \overline{x}_{\pi(n)})]\end{aligned} \tag{A.15}$$

and

$$\begin{aligned}\hat{\mathrm{Agg}}_{\mathrm{Sug}}(\hat{x}_1, \hat{x}_2, ..., \hat{x}_n) = [&\mathrm{Agg}_{\mathrm{Sug}}(\underline{x}_{\pi(1)}, \underline{x}_{\pi(2)}, ..., \underline{x}_{\pi(n)}), \\ &\mathrm{Agg}_{\mathrm{Sug}}(\overline{x}_{\pi(1)}, \overline{x}_{\pi(2)}, ..., \overline{x}_{\pi(n)})] \, .\end{aligned} \tag{A.16}$$

Interval triangular norms and conorms. This approach can also be used to obtain interval aggregation operators based on triangular operations

$$\hat{\mathrm{Agg}}_{\Phi}(\hat{x}_1, \hat{x}_2, ..., \hat{x}_n) = [\mathrm{Agg}_{\Phi}(\underline{x}_1, \underline{x}_2, ..., \underline{x}_n), \mathrm{Agg}_{\Phi}(\overline{x}_1, \overline{x}_2, ..., \overline{x}_n)] \ . \tag{A.17}$$

Appendix A.2. Thresholding strategies

Thresholding is the third step in the proposed approach, which has the aim of converting a numerical or interval decision into a final diagnosis. This subsection lists all implemented and evaluated strategies for both numerical and interval modes.

Appendix A.2.1. Numerical mode

For numerical decisions there is only one class of thresholding strategies – thresholding with margin $\epsilon \in [-0.5, 0.5]$ – given by

$$\tau_\epsilon(a) = \begin{cases} \text{B} & \text{if } a < 0.5 - \epsilon \\ \text{M} & \text{if } a \geq 0.5 + \epsilon \\ \text{NA} & \text{otherwise}\,. \end{cases} \tag{A.18}$$

Appendix A.2.2. Interval mode

For interval mode we evaluated three thresholding strategies. The first approach is to apply a numerical threshold to the interval representative, which results in

$$\hat{\tau}_{\text{rep},\epsilon}([a,b]) = \tau_\epsilon(\text{rep}([a,b]))\,. \tag{A.19}$$

The second is the interval version of thresholding with a margin given for each $\epsilon \in [-0.5, 0.5]$ by

$$\hat{\tau}_\epsilon([a,b]) = \begin{cases} \text{B} & \text{if } b < 0.5 + \epsilon \\ \text{M} & \text{if } a \geq 0.5 - \epsilon \\ \text{NA} & \text{otherwise}\,. \end{cases} \tag{A.20}$$

The last approach involves calculation of the common part between intervals. Let $len([a,b])$ denote the length of interval $[a,b]$. Then this thresholding strategy is given by

$$\hat{\tau}_{\text{cp}}([a,b]) \begin{cases} \text{NA} & \text{if } |[a,b] \cap [0.5-\epsilon, 0.5+\epsilon]| > \\ & \quad \max(|[a,b] \cap [0.5+\epsilon, 1]|, |[a,b] \cap [0, 0.5-\epsilon])| \\ \text{B} & \text{if } |[a,b] \cap [0.5+\epsilon, 1]| < |[a,b] \cap [0, 0.5-\epsilon]| \\ \text{M} & \text{otherwise}\,. \end{cases} \tag{A.21}$$

Acknowledgments

The project was supported by the Microsoft Research Award and the Innovation Incubator program of the Polish Ministry of Higher Education.

[1] A. Wójtowicz, P. Żywica, et al., Dealing with Uncertainty in Ovarian Tumor Diagnosis, in: K. Atanassov, W. Homenda, et al. (Eds.), Modern Approaches in Fuzzy Sets, Intuitionistic Fuzzy Sets, Generalized Nets and Related Topics. Volume II: Applications, SRI PAS, Warsaw, 2014, pp. 151–158.

[2] A. Stachowiak, K. Dyczkowski, et al., A Bipolar View on Medical Diagnosis in OvaExpert System, in: T. Andreasen, H. Christiansen, et al. (Eds.), Flexible Query Answering Systems 2015: Proceedings of the 11th International Conference FQAS 2015, Cracow, Poland, October 26-28, 2015, Springer, 2016, pp. 483–492.

[3] T. De Waal, J. Pannekoek, S. Scholtus, Handbook of statistical data editing and imputation, Vol. 563, John Wiley & Sons, 2011.

[4] World Health Organization, Mortality database, `http://www.who.int/healthinfo/mortality_data/`, online. Accessed on 17/11/2014.

[5] M. Stukan, M. Dudziak, et al., Usefulness of diagnostic indices comprising clinical, sonographic, and biomarker data for discriminating benign from malignant ovarian masses, Journal of Ultrasound in Medicine 34 (2) (2015) 207–217.

[6] R. Moszyński, P. Żywica, et al., Menopausal status strongly influences the utility of predictive models in differential diagnosis of ovarian tumors: An external validation of selected diagnostic tools, Ginekologia Polska 85 (12) (2014) 892–899.

[7] B. Van Calster, K. Van Hoorde, et al., Evaluating the risk of ovarian cancer before surgery using the ADNEX model to differentiate between benign, borderline, early and advanced stage invasive, and secondary metastatic tumours: prospective multicentre diagnostic study, British Medical Journal 349 (2014) g5920.

[8] L. Zannoni, L. Savelli, et al., Intra- and interobserver agreement with regard to describing adnexal masses using International Ovarian Tumor Analysis terminology: reproducibility study involving seven observers, Ultrasound in Obstetrics & Gynecology 44 (1) (2014) 100–108.

[9] S. Hatch, Snowball in a Blizzard: A Physician's Notes on Uncertainty in Medicine, Basic Books, New York, 2016.

[10] P. Żywica, A. Wójtowicz, et al., Improving medical decisions under incomplete data using interval–valued fuzzy aggregation, in: Proceedings of 9th European Society for Fuzzy Logic and Technology (EUSFLAT), Gijón, Spain, 2015, pp. 577–584.

[11] R. E. Moore, Interval analysis, Vol. 4, Prentice-Hall Englewood Cliffs, 1966.

[12] G. Beliakov, A. Pradera, et al., Aggregation functions: A guide for practitioners, Springer, Berlin Heidelberg, 2007.

[13] G. Deschrijver, E. Kerre, Aggregation operators in interval-valued fuzzy and atanassov's intuitionistic fuzzy set theory, in: H. Bustince, J. Herrera, Fand Montero (Eds.), Fuzzy Sets and Their Extensions: Representation, Aggregation and Models, Springer, 2008, pp. 183–203.

[14] R. Yager, OWA aggregation of intuitionistic fuzzy sets, International Journal of General Systems 38 (6) (2009) 617–641.

[15] G. Deschrijver, E. Kerre, Implicators based on binary aggregation operators in interval-valued fuzzy set theory, Fuzzy Sets and Systems 153 (2) (2005) 229–248.

[16] G. Beliakov, H. Bustince, et al., Aggregation for Atanassov's intuitionistic and interval valued fuzzy sets: The median operator, IEEE Transactions on Fuzzy Systems 20 (3) (2012) 487–498.

[17] D. Dubois, H. Prade, Formal representations of uncertainty, Decision-making Process: Concepts and Methods (2009) 85–156.

[18] D. Szpurek, R. Moszyński, et al., An ultrasonographic morphological index for prediction of ovarian tumor malignancy, European Journal of Gynaecological Oncology 26 (1) (2005) 51–54.

[19] J. L. Alcázar, L. T. Mercé, et al., A new scoring system to differentiate benign from malignant adnexal masses, Obstetrical & Gynecological Survey 58 (7) (2003) 462–463.

[20] D. Timmerman, A. C. Testa, et al., Logistic regression model to distinguish between the benign and malignant adnexal mass before surgery: a multicenter study by the International Ovarian Tumor Analysis Group, Journal of Clinical Oncology 23 (34) (2005) 8794–8801.

[21] D. Timmerman, T. H. Bourne, et al., A comparison of methods for preoperative discrimination between malignant and benign adnexal masses: the development of a new logistic regression model, American Journal of Obstetrics and Gynecology 181 (1) (1999) 57–65.

[22] I. Jacobs, D. Oram, et al., A risk of malignancy index incorporating CA 125, ultrasound and menopausal status for the accurate preoperative diagnosis of ovarian cancer, BJOG: An International Journal of Obstetrics & Gynaecology 97 (10) (1990) 922–929.

[23] N. Japkowicz, M. Shah, Evaluating Learning Algorithms: A Classification Perspective, Cambridge University Press, New York, NY, USA, 2011.

[24] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria (2014). URL http://www.R-project.org

[25] R. R. Yager, On ordered weighted averaging aggregation operators in multicriteria decisionmaking, IEEE Transactions on Systems, Man and Cybernetics 18 (1) (1988) 183–190.

[26] R. Mesiar, A. Stupňanová, R. R. Yager, Generalizations of OWA operators, IEEE Transactions on Fuzzy Systems 23 (6) (2015) 2154–2162.