Computers in Biology and Medicine 202 (2026) 111416

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Weakly supervised treatment selection: Machine learning models for appropriate surgical planning of submandibular stones

Andrea Campagner[a,b,1], Matteo Lazzeroni[c,d,1], Caterina Pizzi[e], Caterina Sattin[e], Giulia Buccichini[f], Massimo Del Fabbro[d], Gianluca Martino Tartaglia[d,i], Maria Cristina Firetto[e], Gianpaolo Carrafiello[e], Michael Koch[g], Pasquale Capaccio[d,h,*], Federico Cabitza[a,b,**]

[a] Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milan, Italy
[b] IRCCS Ospedale Galeazzi Sant'Ambrogio, Milan, Italy
[c] Department of Otorhinolaryngology & Head and Neck Surgery, Amsterdam UMC, University of Amsterdam, Amsterdam, Netherlands
[d] Department of Biomedical, Surgical and Dental Sciences, University of Milano, Milan, Italy
[e] Operative Unit of Radiology, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico Di Milano, Milan, Italy
[f] Department of Otolaryngology and Head and Neck Surgery, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico di Milano, Milan, Italy
[g] Department of Otolaryngology and Head and Neck Surgery, University of Erlangen-Nuremberg, Erlangen, Germany
[h] Department of Otolaryngology and Head and Neck Surgery, Fatebenefratelli Hospital, ASST Fatebenefratelli Sacco, Milan, Italy
[i] Unit of Maxillofacial Surgery and Dentistry, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy

HIGHLIGHTS

- We propose a machine learning model to support surgical planning for salivary stones.
- The model uses weakly supervised learning to handle uncertain treatment outcomes.
- We apply credal learning to encode clinical uncertainty using probability sets.
- Our method outperforms traditional and causal models in treatment prediction.
- Radiological features like stone distance and volume emerge as key predictors.
- We introduce a practical, interpretable tool to guide personalized treatment choices.

ARTICLE INFO

Keywords:
Medical machine learning
Treatment selection
Weakly supervised learning
Credal learning
Sialolithiasis
Explainable AI

ABSTRACT

There is a gap in real-world clinical adoption of machine learning (ML) solutions due to the inherent uncertainty and variability in treatment outcomes. To bridge this gap, we present a novel approach to the problem of medical treatment selection using ML models and we apply it to the case of submandibular sialolithiasis treatment. The study introduces a weakly supervised learning framework which allows for the inclusion of imprecise, incomplete, or noisy ground truth data. By applying this methodology to the specific medical problem of submandibular stone treatment, we demonstrate the potential of encoding treatment outcomes as credal sets—collections of probability distributions reflecting the uncertain nature of the optimal treatment—to improve surgical planning and decision-making. We validated our model using real-world patient data, showcasing its ability to offer personalized treatment recommendations based on radiological features of submandibular stones. Our study underscores the importance of incorporating proper uncertainty management into ML for clinical practice to support clinical decision-making, by showing a promising solution to improve the treatment of sialolithiasis.

* Corresponding author at: Department of Biomedical, Surgical and Dental Sciences, University of Milano, Milan, Italy.
** Corresponding author at: Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milan, Italy.
Email addresses: pasquale.capaccio@unimi.it (P. Capaccio), federico.cabitza@unimib.it (F. Cabitza).

[1] Authors contributed equally

https://doi.org/10.1016/j.compbiomed.2025.111416
Received 18 December 2024; Received in revised form 17 December 2025; Accepted 18 December 2025
Available online 2 January 2026
0010-4825/© 2025 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Surgical planning, or treatment selection, is one of the most fundamental problems in clinical decision- making. Once a diagnostic hypothesis is formulated, identifying the most appropriate treatment course becomes a crucial challenge; clinicians typically rely on multiple information sources, including their own experience, existing evidence, established guidelines, or routine practice, to address this challenge. In recent years, Artificial Intelligence (AI) methods, and Machine Learning (ML) in particular, have been proposed to support treatment selection [27,31], enabling recommendations for the most appropriate treatment to be grounded in large volumes of multivariate data, which would be impossible to analyze systematically using traditional means.

Yet, despite promising results in simulated and laboratory settings, ML-based solutions for treatment selection have yet to be widely adopted in real-world clinical practice [7,47]. A primary reason for this gap is that the treatment selection problem is fraught with uncertainty, arguably even more so than other healthcare tasks such as diagnosis and prognosis. Indeed, not only does one face the problem of dealing with variability and measurement errors [10], which are ubiquitous in the life sciences, but also the very "quality" of the ground truth for treatment selection can be called into question. Indeed, one is never able to observe the optimal treatment course, but rather only the one that has actually been performed and its outcome [7,17].

To better understand this issue consider the following scenario: a patient presents to a clinician, who formulates a diagnosis D: a condition for which there are two possible treatment procedures, namely treatment A and B. Based on the patient's characteristics and following evidence-supported medical guidelines, the doctor selects A as the treatment: Initially, the patient's condition improves, however, they develop a post-treatment complication and unfortunately die. Crucially, we have no way to determine whether treatment A was actually the optimal treatment for the patient's condition. Perhaps, if treatment B had been performed instead, the patient would have fully recovered; by contrast, the patient death may be due to another condition and they would have died the same even if treatment B had been performed. The key issue is that treatment B has never been performed, and thus we have no way to determine whether it would have been the optimal treatment option: we only know that the choice of treatment A was supported by some evidence.

This issue, that is the lack of an objective ground truth, is an intrinsic characteristic of the treatment selection problem. Ideally, solving this problem would require information regarding all *counterfactual scenarios* [41]; however, the inaccessibility of such data renders this task significantly more complex than other clinical decision-making challenges, thus limiting the adoption of ML-based solutions in clinical practice. Indeed, even though ML techniques aimed at addressing this issue have been proposed [31,36,41], such techniques require strong statistical assumptions, such as the ability to infer a fully specified causal model, which are typically difficult, if not impossible, to verify [19,38] or enforce in practical contexts. Without these assumptions, there are no guarantees regarding the method's ability to infer optimal treatments [6].

In this article, we propose a novel, computationally lightweight approach to address the treatment selection problem based on the idea of acknowledging the intrinsic limitation in ground truth quality and recasting this problem as a weakly supervised learning task. Weakly supervised learning [57] (WSL) refers to ML tasks in which the ground truth is far from the perfect idealization often assumed in supervised learning, but could instead be incomplete [53], noisy [45] or imprecise [25]. In this setting, we employ weak supervision as a way to encode the information that the treatment supervision we observe does not represent a fully certain indication of the optimal treatment, but only a piece of evidence that the treatment may be better suited than others, which cannot, nevertheless, be fully excluded from further consideration. We will focus specifically on the WSL framework called credal learning [11]: compared with other approaches [14], credal learning has attracted increasing interest in the recent years due to its ability to model different ground truth quality issues (including errors and imprecision). In credal learning, the weak ground truth information is encoded in terms of a credal set [4]: that is, a collection of probability distributions that express partial knowledge or uncertainty about the true label. In the context of treatment selection, such credal sets encode prior beliefs about the true, but unknown, value of the ground truth that we seek to recover. They represent all available information regarding the treatment administered to each patient in the training set in the form of a set of probabilities that reflects the available evidence, which is weak and counterfactually incomplete.

Grounding on this WSL approach, in this article we make the following contributions: first, we recast the treatment selection problem as a WSL task using the credal learning framework, showing conditions under which, despite the above mentioned ground truth quality issue, the optimal treatment selection can nevertheless be recovered; second, we will demonstrate the effectiveness of our approach by means of an application on a real-world medical problem, namely the problem of treatment selection in the context of sialolithiasis, a common condition that clinically presents with recurrent swelling and pain of the salivary glands [40]. To this end, we aimed at answering two research questions: RQ1) Does the use of WSL lead to an improvement in performance (i.e., accuracy) as compared to traditional ML techniques for treatment selection in sialolithiasis? RQ2) Does the proposed approach provide sufficiently reliable recommendations and could it be used to provide practical, accessible tools for physicians to select the correct therapeutic strategy for submandibular stones while relying on information from radiological features?

The rest of this article will be structured as follows. In Section 2 we present the methodology of our study. Specifically, after introducing the relevant background in Sections 2.1 and 2.2, we introduce our methodology for WSL in Section 2.3, where we also analyze its properties from the point of view of learning theory. Then, in Section 2.4 we present the case study to which we applied our WSL methodology. In Section 3 we present the results of our empirical investigation, which we then discuss in Section 4. Finally, in Section 5, we summarize our contributions and conclude the article.

2. Methods

2.1. Formal notation and supervised learning

Let X be a vector space in $\mathbb{R}^d$, which we assume to be the space of features, representing the set of characteristics of patients based on which we aim to decide the appropriate treatment course. Let Y be a finite set of labels, representing the set of all admissible treatments. We assume a joint distribution P defined over $X \times Y$, which defines the data generating process, associating patients with the corresponding optimal treatment assignment. Let H be a class of models, and let $l : X \times Y \times H \rightarrow R$ be a loss function, where $l(x, y, h)$ evaluates the cost of applying treatment $h(x)$ when the optimal treatment is y.

The main object of interest is the conditional distribution $P(\cdot|x)$ which, given a patient's feature representation x, describes the probability that a given treatment is the optimal one. The aim of supervised learning is, given a finite sample $S = (x_i, y_i)_{i=1}^N$, to find a model $h \in H$ that minimizes the true risk $R_P(h) = \mathbb{E}_P[l(x, y, h)] = \int l(x, y, h)P(x, y)$. However, since the distribution P is unknown (otherwise the problem would be trivially solvable by simply setting h to be the Bayes predictor), one is restricted to consider heuristic criteria for model selection, such as minimizing the empirical risk $R_S(h) = \frac{1}{N}\sum l(x_i, y_i, h)$ [44]). Results from statistical learning theory ensure that, under reasonable assumptions, empirical risk provides a good approximation of the true risk [54].

The previous formalization, while widely employed in practice, assumes that the true supervision for any instance x, in our case, the true

optimal treatment, is observable. According to the Bayesian interpretation of how decision-making operates in medical practice [3,24,51], however, given a patient x, a clinician will select the treatment based on a distribution $L(\cdot|x)$ that describes the clinician's pre-existing knowledge (as well as, possibly, biases) and available evidence: crucially, $L(\cdot|x)$ could be different from the true conditional distribution $P(\cdot|x)$. As the performed treatment is drawn from $L(\cdot|x)$, rather than $P(\cdot|x)$, any ML models trained using the observed treatment labels will optimize the loss function $R_L(h) = \int l(x, y, h)L(y|x)P(x)$, which could, however, be significantly different from the true risk $R_P(h)$. To avoid the risk of learning a model that may overfit on L (thus, may be arbitrarily bad for the true conditional distribution P), as hinted at in the Introduction, we formalize treatment selection as a WSL problem: in particular, a credal learning problem [11].

2.2. Background on weakly supervised learning

In the setting of WSL, the real label y corresponding to a case representation x is not observable. Instead, we are only able to observe an imprecise, possibly incorrect, label. In the context of credal learning [11,34], this imprecise version of y takes the form of a credal set $C(x)$, which is a convex, closed set of probability distributions over Y. Such a credal set has an epistemic interpretation [42], in that it represents the belief (held by the annotator, i.e., in the setting of treatment selection, the clinician) that the true conditional distribution $P(\cdot|x)$ lies in $C(x)$. Notably, however, $P(\cdot|x)$ may not belong to $C(x)$. By contrast, we will assume that the (potentially wrong) probability distribution $L(\cdot|x)$ (i.e., the distribution by which the clinician selects the treatment to be performed and, consequently, from which the observed label is drawn) is contained in $C(x)$.

We assume data to be generated from a joint distribution M defined over $X \times Y \times Q(Y)$, where $Q(Y)$ is the collection of all credal sets over Y such that $L(\cdot|x) \in Q(x)$. In particular, we assume data is sampled from $M \downarrow X \times Q(Y)$, that is, the projection of M onto $X \times Q(Y)$; therefore, we assume that the true labels y are not directly observed. Intuitively, thus, the observed data can be understood to be generated according to the following two-step procedure: first an instance (x, C, y), where C is a credal set and y is the true label, is sampled from M, then y is discarded and only (x, C) is observed.

As in the setting of supervised learning, our aim is to find a model h that minimizes the true risk with respect to the true labels. However, in contrast to supervised learning, in the setting of credal learning not only is the data generating distribution unknown, but also the true labels y are not observed: therefore, empirical risk minimization cannot be applied [9]. To generalize empirical risk minimization to the setting of credal learning one can consider a generalized loss function [11,25,34] which lifts a loss function l to a loss function $l_Q : X \times Q \times H \to R$ that evaluates the output of the model h against the observed credal sets in the training set. While such an approach can be theoretically justified, and indeed leads to similar theoretical guarantees as for the case of supervised learning [8,9], in this article we will not describe such an approach due to the fact that its full generality (with the associated limitations, in particular related to computational complexity) is not necessary for the setting of treatment selection we consider in this article: we refer the interested reader to [11].

2.3. A weakly supervised formalization of treatment selection

As noted in the Introduction, the treatment selection problem can be naturally framed as a WSL task. Indeed, the fact that for a given patient x we have observed that a given treatment y has been performed does not provide fully conclusive evidence that y is indeed the optimal treatment course, due to the fact that, as described in the previous Section, $L(\cdot|x)$ will in general be different from $P(\cdot|x)$. Thus, even though the goal of the treatment selection problem is to learn the distribution $P(y|x)$, i.e., the probability that y is the optimal treatment for patient x, we are unable to observe such distribution: instead, we are only able to observe the treatment y (selected according to distribution $L(\cdot|x) \neq P(\cdot|x)$) that has been performed on x.

In analogy with the epistemic semantics of credal sets described above, such a treatment indication y can be understood as a description of the epistemic state of the clinician concerning $P(\cdot|x)$. Notably, this epistemic state is dynamic, in the sense that it can change during and after surgery, based on the available information and evidence at the time. Indeed, both during and after the surgery, there can be complications or the need for post-failure re-intervention through another treatment (which would have been more appropriate in the first place) that may change the epistemic state of the clinician. Crucially, such information is available during model training, since this latter is usually performed on retrospective data [27]. Therefore, based on the outcome of y, we can infer some constraints [43] about the epistemic state of the clinician. In particular, we can distinguish three different outcomes for a given treatment y, each of which corresponds to a different epistemic state held by the clinician:

1. Treatment y has been performed and no adverse events have been observed. As mentioned above, the fact that y has been performed suggests that the clinicians believed that y could have been the optimal treatment. Hence, the evidence in favor of y being the optimal treatment is larger than for all other possible alternative treatments: that is, the clinician held the belief that $P(y|x) > P(y'|x)$ for all other $y' \in Y$;
2. Treatment y was initially planned for, but due to a failure (e.g., caused by the emergence of some condition that was not known or foreseeable at the time of the initial treatment planning) a different treatment y' is ultimately performed. In this case, the evidence suggests that treatment y' could have been preferred to y from the start (thus avoiding the failure), that is $P(y'|x) > P(y|x)$ (notice, that this is an a posteriori epistemic state; that is, the epistemic state after observing the failure). At the same time, the fact that initially treatment y had been selected, suggests that the evidence in favor of y being the optimal treatment was larger than the evidence for all other alternatives (except, a posteriori, y'), thus $P(y'|x) > P(y|x) > P(y''|x)$ for all $y'' \in Y$ with $y'' \notin \{y, y'\}$;
3. Finally, we have the situation in which y has been performed but it ultimately resulted in an adverse event: adverse events may include complications, death, or any other worsening of the patient's condition or quality of life. In this case, even though the initial evidence pointed in the direction of treatment y, we have no information about the optimal treatment, as we do not know what would have happened if any other treatment had been performed (that is, whether the adverse event would have occurred the same or not). Thus, a posteriori, the epistemic state of the clinician can be encoded as the vacuous epistemic state "?", in which they have no information at all about $P(\cdot|x)$.

The technical justification for our approach lies in the observation that each of the epistemic states identified above can be represented as a credal set endowed with a special structure. Indeed, the following result (together with Corollary 1 in the Appendix) shows that the epistemic states described above can be modeled as credal sets.

Theorem 1. *Let Y be a finite set. Let T be a tree[2] whose nodes are labeled with elements from some subset $Y' \subseteq Y$. We say that a probability distribution P is compatible with T if $(y, y') \in T \implies P(y) \geq P(y')$. Then, the set $Q(T)$ of probability distributions compatible with P is a credal set (that is, it is convex and closed). Furthermore, $Q(T)$ is the set of probability distributions such that:*

[2] A tree T is a pair $V(T), E(T)$, where $V(T)$ is the set of vertices and $E(T) \subseteq V(T) \times V(T)$ is the set of edges. With an abuse of notation, given $v, v' \in V(T)$, we write $(v, v') \in T$ instead of $(v, v') \in E(T)$.

1. $P(y) \leq 1$ *if* y *is the root of* T *or* y *is not in* T*;*
2. $P(y') \leq \frac{1}{d}$*, where* d *is the depth of* y' *in* T*, otherwise.*

Thus, each tree T *can be associated with a normalized possibility distribution [20]* $poss_T : Y \to [0,1]$ *defined by* $poss_T(y) = max_{P \in Q(T)} P(y)$.

Thus, the previous result shows how the epistemic states of a clinician in the treatment selection problem can be represented as credal sets: first, the epistemic state held by the clinician is represented in terms of a tree (as described in Corollary 1 in the Appendix), then this latter is represented as a credal set (constructed according to Theorem 1 and Corollary 1). Consequently, the treatment selection problem can be formalized as a problem of credal learning, where the credal sets associated with each instance take the form of a possibility distribution. Intuitively, this implies that the treatment selection problem could in principle be solved through the following steps:

1. Given a training set $S = (x_i, y_i)$, we transform each label y_i into a credal set by applying Corollary 1, thus obtaining a weakly supervised training set $S_I = (x_i, Q_i)$. Notice that this requires having information about failures or adverse events for the patients in the training set;
2. Apply any learning algorithm for credal learning to obtain a model h from the weakly supervised training set S_I.

We notice that step (1) in the previous procedure can be performed efficiently, indeed, as a consequence of the previous result, for each instance x_i, the possibility distribution $poss_i$ corresponding to the credal set Q_i can be easily constructed in time $\Theta(|Y|)$ (e.g., by applying depth-first search [16]). Thus, if the time complexity required to train a model h in step (2) is also polynomial, then the treatment selection problem can be solved in polynomial time. This is in contrast with the general case of credal learning which, without additional assumptions on the structure of the credal sets, is NP-HARD [11]. Thus, to show that our approach can be effectively applied to solve the treatment selection problem we need to prove two things: (1) step 2 above can be solved efficiently; (2) the above described reduction of treatment selection to credal learning is able to recover the true optimal treatment.

In order to solve the first problem, and thus address step (2), we consider a generalization of the Random Resampling-based Learning (RRL) algorithm, originally proposed in the context of learning from fuzzy labels [9]. The algorithm is described in Algorithm 1. This algorithm is a generalization of RRL to the setting of treatment selection (as a special form of credal learning). Intuitively, RRL constructs an ensemble of models based on different possible versions of the ground truth: each of these versions represents a possible configuration for the optimal treatment. Crucially, while some of these configurations will contain a wrong treatment assignment, on average and unless the clinician's distribution $L(\cdot|x)$ is far from the true conditional distribution $P(\cdot|x)$, we expect them to match the optimal treatment assignment with high probability. Then, by ensembling, RRL smooths away the classification errors and recovers the optimal treatment assignment. In regard to the properties of RRL, it can be easily shown that Algorithm 1 ensures that the total time complexity required to solve the treatment selection problem is polynomial.

Proposition 1. *Let* H *be a base hypothesis class. Let* A *be a polynomial-time learning algorithm for* H*. Then, Algorithm 1 has time complexity* $O(k * T_A(|S|) + |S||Y|)$*, where* $T_A(|S|)$ *is the time complexity of* A *on a dataset of size* $|S|$.

The above result, together with Theorem 1 and Corollary 1, demonstrates that treatment selection within the proposed credal learning framework can be performed efficiently—in polynomial time. As previously noted, this stands in sharp contrast to other credal learning approaches, which in the worst case require solving NP-HARD optimization problems [11]. It also contrasts with earlier treatment selection methods based on Bayesian frameworks, which often rely on approximate inference techniques—such as Monte Carlo simulations—to circumvent computational complexity [46].

Algorithm 1 The RRL algorithm for the treatment selection problem.

procedure RRL($S = (x, y, m)$: dataset (x: features, y: target, m: information about complications and failures), k: ensemble size, $\mathcal{H}$: model class)
  $Ensemble \leftarrow \emptyset$
  $\tilde{S} \leftarrow \emptyset$
  for all $(x, y, m) \in S$ **do**
    if m = complication **then**
      $T \leftarrow$ "?"
    else if m = failure **then**
      $T \leftarrow m_y > y > y'$, $\forall y' \in Y \rhd m_y$ is the post-failure treatment
    else
      $T \leftarrow y > y'$, $\forall y' \in Y$
    end if
    Construct $Q(T)$
    $\tilde{S}$.append($(x, Q(T))$)
  end for
  for all iterations $i = 1$ to k **do**
    Draw a boostrap sample S' from $\tilde{S}$
    $Tr_i \leftarrow \emptyset$
    for all $(x, \pi) \in S'$ **do**
      Sample $\alpha \sim Uniform[0,1]$
      Add (x, y') to Tr_i, where $y' \sim Uniform[\{y \in Y : poss_T(y) \geq \alpha\}]$
    end for
    $h_i \leftarrow \arg\min_{h \in \mathcal{H}} L_{Tr_i}(h) = \frac{1}{|Tr_i|} \sum_{(x,y') \in Tr_i} l(x, y', h)$
    Add base model h_i to $Ensemble$ with weight $\frac{1}{L_{Tr_i}(h)}$
  end for
  return $Ensemble$
end procedure

More relevantly, we show that Algorithm 1 is actually able to solve the treatment selection problem, providing generalization guarantees on the recovery of the possibly unobserved optimal treatment. In particular, the following result shows that, as long as the credal set over treatment alternatives constructed by applying Corollary 1 contains the true probability distribution for the optimal treatment, i.e., $P(\cdot|x)$, RRL recovers the true optimal treatment.

Theorem 2. *Assume that, with probability 1 over the sampling of an imprecise instance* $(x, y, Q(x))$ *drawn from* M*, where* $Q(x)$ *is constructed as in Corollary 1, it holds that* $P(\cdot|x) \in Q(x)$*. Assume, further, that* P *is realizable for* $\mathcal{H}$ *(i.e., there exists* $h \in \mathcal{H}$ *s.t. with probability 1* $h(x) = P(\cdot|x)$*). Then, asymptotically (i.e., when the sample size* n *and ensemble size* k *go to infinity), it holds that* $L_P(RRL) = min_f L_P(f^*)$*, where* f *ranges over the functions measurable with respect to* P*. That is, asymptotically, RRL converges to a Bayes predictor.*

The previous result shows that, if the belief of the clinician is closely aligned with the true optimal treatment probability $P(\cdot|x)$ (that is, the epistemic states $Q(x)$, constructed from the observed treatments that have been performed, always contain $P(\cdot|x)$), then RRL, in the limit of infinite data, would be able to recover the optimal treatment selection. In Appendix B we evaluate the robustness of RRL to violations of the assumption. However, Theorem 2 holds only asymptotically and does not provide an explicit bound on the true risk, $L_P(RRL)$, as a function of the sample size or the training error. The following theorem, then, strengthens Theorem 2, showing that the error rate of RRL decreases exponentially fast with respect to the ensemble size and the validation error of the worst model in the ensemble:

Theorem 3. *Assume RRL is executed with an ensemble of size k. Assume, further, that the conditions for Theorem 2 hold. Then, assuming that with probability 1 the error rates of the models sampled using RRL are independent, for each $\delta > 0$, with probability greater than $1 - \delta$, it holds that:*

$$L_P(RRL) \leq e^{-m \ KL(0.5||g_V)}, \tag{1}$$

where KL is the Kullback-Leibler divergence [37] and

$$g_V = max_{h \in H_A} OOB(h) + \sqrt{\frac{\log(2n/\delta)}{n_{OOB(h)}}} \leq 0.5,$$

where $OOB(h)$ is the out-of-bag error of model h and $n_{OOB(h)}$ is the size of the corresponding out-of-bag validation set (that is, the set of instances that have not been used to train h).

Proof. The result follows from Theorem 3.6 in [9]. □

Finally, we show that, under reasonable assumptions, the RRL algorithm is guaranteed (with high probability) to have lower error than simply employing the (potentially incorrect) labels drawn from $L(\cdot|x)$: that is training RRL from the credal sets $Q(x)$ constructed from the observed treatment labels drawn from $L(\cdot|x)$ is expected to provide better results than training a traditional model directly based on the same labels.

Theorem 4. *Assume the conditions for Theorem 3 hold. Let*

$$\eta = \mathbb{E}[KL(P(\cdot|x)||L(\cdot|x))] > 0,$$

where $P(\cdot|x)$ is the true optimal treatment distribution for instance x, $L(\cdot|x)$ is the clinician's treatment assignment distribution for instance x, and the expectation is w.r.t. $M \downarrow X$ (the marginal of M w.r.t. X). Let A be a learning algorithm for base class H. Then, it holds that

$$\mathbb{E}[L_P(A) - L_P(RRL)] \in \Theta(\eta),$$

where the expectation is w.r.t. the sampling of a training set and the sampling of models in RRL.

Proof. The result follows from Theorem 2.1 in [11], noting that the excess risk of $L_P(A)$, as compared with $L_P(RRL)$, is asymptotically equal to η. □

Using the previous theorems, we have proved that, under reasonable assumptions, the proposed methodology can recover the (unknown) optimal treatment with a higher probability than relying solely on the available (noisy) labels. In the following sections, as mentioned in the Introduction, we will describe an experiment in which we evaluated the proposed approach in the setting of treatment selection for sialoadenectomy, in comparison with both traditional and state-of-the-art approaches to treatment selection. Before proceeding, however, it is important to highlight the potential limitations of the proposed approach. First, the credal learning formalization of treatment selection requires that information about treatment outcomes (e.g., complications, surgical failures, or other adverse events) be available as a proxy for the (sub-)optimality of the treatment received by patients. While such information is typically available in training data, its absence would render the proposed method inapplicable—or, more precisely, equivalent to training on the available labels.

A second limitation concerns the technical conditions under which the method is guaranteed (with high probability) to recover the optimal treatment. Specifically, according to Theorem 2, a sufficient condition for recoverability is that the true probability distribution of the optimal treatment, i.e., $P(\cdot|x)$, always lies within the credal sets constructed from the observed treatments. Common scenarios in which this assumption may fail include pervasive clinician bias, recording errors, or unobserved latent confounders, all of which can introduce a systematic deviation between the observed and optimal treatments—i.e., cases where the distribution $L(\cdot|x)$ differs substantially from $P(\cdot|x)$.

When this assumption does not hold, Theorem 2.1 in [11] implies that recoverability of the optimal treatment cannot be guaranteed. In particular, if the distance between the true probability $P(\cdot|x)$ and the credal set $Q(T)$ is γ, then the error rate can be upper-bounded by a term of order $O(\gamma)$.

Nonetheless, this limitation may be less restrictive than it initially appears. On the one hand, since $L(\cdot|x) \in Q(T)$, even when the assumption fails, the proposed method is still guaranteed (with high probability) to achieve a lower asymptotic error rate than simply using the available noisy labels, as established by Theorem 4 (given that $\gamma \leq \eta$). On the other hand, the condition that $P(\cdot|x) \in Q(T)$ is a sufficient but not necessary condition for recoverability. This means that, in principle, recoverability is not ruled out even when the assumption is violated.[3] For example, although no formal results are known for the RRL algorithm, it is known that gradient-based algorithms—such as those originally studied in [11]—are, under certain conditions, robust to noise [1] which, in turn, implies that the optimal treatment may still be recoverable, even when the sufficient condition is not satisfied. We leave to future work a detailed analysis of the failure conditions of the proposed method, and, in particular, the derivation of lower bounds for credal learning.

More generally, several strategies can be employed to enhance the robustness of the proposed method to violations of the aforementioned assumption, thereby mitigating the risk of systematic bias in treatment predictions. These include: dynamically enlarging credal sets based on new outcome data to capture potential deviations in the clinician's epistemic stance from the optimal treatment; clinician review of low-confidence predictions to assess uncertain outputs generated by the RRL model; and flagging epistemically vacuous recommendations for second opinions, thereby incorporating the expertise of multiple clinicians to construct more informative credal sets.

2.4. Case study

2.4.1. Background on sialolithiasis

Sialolithiasis is a condition that clinically presents with recurrent swelling and pain of the affected salivary gland [40], and accounts for approximately 60 %–70 % of obstructive salivary gland disease cases [21]. The stones are generally unilateral and mainly affect the excretory duct system of the submandibular gland (80 %) [12], probably due to the acute angles described by Wharton's duct along its course and to the viscosity of its saliva [30]. A recent work by Schapher et al. [48] has offered important and innovative insight into the aetiopathogenesis of sialoliths; highlighting key factors in its development. According to their results different conditions previously described in literature [39,49] are able to induce an inflammatory reaction, favoring an accumulation in the salivary ducts of neutrophils, ultimately resulting in the gradual development and appositional growth of salivary stones, which eventually obstruct the excretory ducts of the gland.

Diagnosis can be achieved by various radiological investigative techniques, including sonography, computed tomography (CT) and cone beam computed tomography (CBCT). While which of these should be considered the gold standard has been debated [22], all of these diagnostic techniques aim at precisely determining not only the location of the sialoliths, but also their morphological and material characteristics: these can greatly determine the most appropriate treatment course and have generally been associated in the past with varying success rates of sialoendoscopic procedures [35].

Over the last decades treatment of submandibular stones has radically changed [28]. Gland resection, the most invasive option, now plays only a marginal role in the therapeutic algorithm for submandibular

[3] In Appendix B we illustrate in a simulated experiment a setting where recoverability is indeed possible despite the mentioned assumption not being satisfied.

sialoliths [5]. It is now reserved for those calculi that are inaccessible to minimally invasive endoscopic techniques, or in cases of multiple stones and significant glandular inflammation [5,28]. Indeed, the advent of minimally invasive techniques, such as transoral ductal surgery, has significantly changed the state of art for the management of this condition [28], now focused on the functional preservation of the gland. Nonetheless, sialolithiasis management is complicated by the fact that existing guidelines do not adequately account for modern diagnostic approaches, such as Cone Beam Computed Tomography (CBCT), with the consequence that surgeons may need to convert intraoperatively from a minimally invasive approach to traditional surgical options [5,28].

While ML has been applied to oncological diagnosis [26], to the best of our knowledge, there are no prior applications for treatment selection in sialolithiasis.

For these reasons, we considered the application of the proposed methodology in this specific setting.

2.4.2. Patient selection and data acquisition

This study included a cohort of 132 patients with a history of impacted submandibular stones, treated via endoral/transoral resection or sialadenectomy at the ENT department of Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico of Milan between January 2013 and January 2023. All patients were referred for recurrent episodes of painful swelling of the affected gland. Only patients who had undergone a CT (SOMATOM Definition - Siemens Healthcare, Erlangen, Germany) or CBCT (CBCT 3D CS 9300 Carestream, Rochester, New York, USA) examination at the radiological department of the same institution were included. Patients who were treated by means of other techniques were not included in the present study. All the surgical procedures were performed by the same expert ENT specialist (PC), following standard guidelines. This study was conducted in accordance with the Declaration of Helsinki and approved by the local Ethics Committee.

For each patient, the clinical and radiological parameters were acquired, following segmentation (see Section 2.4.3) and subsequently recorded in an electronic database (Excel 2016 v16.0, Microsoft Corporation, Redmond, WA). The selected treatment, along with any surgical complications or treatment failures, was recorded in the same database. The full set of considered parameters encompassed: Age at the time of the radiological investigation (years); Side of the affected gland (left/right); Maximum diameter of the stone (millimetres - mm); Shape of the stone (elongated/oval); Volume of the stone (mm3), obtained from 3D reconstructions of the stone which were acquired through manual segmentation techniques; Density of the stone (HU); Distance from the mandibular symphysis (mm), defined as the length of the line passing through the mandibular symphysis and intersecting a perpendicular line passing through the most anterior point of the stone; Distance from the mandibular body (mm), defined as the length of the line perpendicular to the distance from the mandibular symphysis and enclosed between the most lateral point of the stone and the body of the mandible; Depth of the stone from the mandibular body (mm), defined as the distance from the inferior margin of the mandible to the most superior point of the stone on a coronal plane; Type of surgery (endoral asportation, transoral asportation, sialadenectomy); Removal of the stone in one piece or multiple fragments; Failure of the surgical treatment; Complications of surgery (ductal stenosis, ranula, postoperative bleeding, infection, lingual nerve injury, recurrence of submandibular stones).

2.4.3. Segmentation

Stone segmentation was performed in order to obtain the volume and other radiological features of each submandibular gland stone. Segmentation was performed through 3D Slicer 4.0 semi-automatic segmentation system, which is freely downloadable from the website http://www.slicer.org. The following workflow was used:

1. Loading of the Digital Imaging and Communications in Medicine (DICOM) data into 3D Slicer;
2. Selection of the 3D multiparametric images and segmentation of the salivary gland stones on the axial images;
3. Selection of the "Segment Editor" module to delineate the salivary stones;
4. Segmentation of the sialoliths with the "Level Tracing" option, by which moving the mouse over the area to be segmented defines an outline where the pixels all have the same background value as the current background pixel, which allows us to segment the chosen area in a more practical and quicker way;
5. Selection in the software's drop-down menu of the option "Quantification" and then "Segment Statistic" to measure the volume of the segmented area thus obtaining the volume of the stone itself.

The hardware platform used was an Apple MacBook Air (1,1 GHz Intel Core i3 dual-core, 8 GB RAM, Intel Iris Plus Graphics 1536 MB, Mac OS 3 12.6 Monterey).

2.4.4. Model design and development

The target task was to predict the optimal treatment option for any given patient. As this target is not observable, we modeled the learning task as a weakly supervised problem, according to the approach presented in Section 2.3.

Specifically, as a pre-processing step, we divided the cases into three categories: (1) cases associated with no failure and no complication; (2) cases associated with a failure of the surgical treatment (and conversion to sialoadenectomy); (3) cases associated with a complication of surgery. Applying the results in Section 2.3, we transformed the labels into credal sets. In particular, we applied the following transformations: for case 1) the original treatment indication T was transformed into the weakly supervised label "$T > A; T > B$", where A and B are the two treatment options that were not applied; for case 2), we transformed the original treatment T into the weakly supervised label "$\text{sialoadenectomy} > T > A$", where A is the third treatment option that was excluded in favor of T; finally, for case 3), we transformed the original treatment T into the vacuous weakly supervised label "?". The above weakly supervised labels were then transformed into credal sets by application of Corollary 1.

No further pre-processing of the data was performed before data splitting. In particular, as there were no missing values, no imputation was performed.

The WSL models were developed based on the RRL algorithm (see Algorithm 1). We considered five base classes of models, namely: Decision Tree (DT), multinomial Logistic Regression (LR), Support Vector Machine (with linear kernel) (SVM), Random Forest (RF), and eXtreme Gradient Boosting (XGB). We selected these base models as they have been shown to offer state-of-the-art performance on tabular data [23,50], as well as for having a varied selection of interpretable (DT, LR, SVM) and black-box (RF, XGB) models.

Due to the limited amount of available data, as well as the requirement of having interpretable models, we fixed the hyper-parameter values of the different models: in particular, to address the label imbalance (see Table 1), we set the $class\_weight$ parameter, for all models, to "balanced", while we set the number of models in the RRL algorithm to 100, the depth of decision trees in the DT, RF and XGB models to 5, and the kernel for the SVM model to linear. For all other hyper-parameters, we set the default values in the `scikit-learn` library. Thus, we did not perform hyper-parameter optimization.

We considered three different evaluation procedures for the ML models.

First, we compared the WSL models against the fully-supervised ones, with the aim of empirically evaluating the results in Section 2.3, and specifically Theorem 4. That is, we assessed, in a real clinical scenario with limited data, whether the RRL algorithm outperforms models trained on the original (potentially incorrect) treatment assignments. To this aim, the fully-supervised models were trained using the

Table 1
Distribution of the features, in terms of mean and standard deviation (for the continuous features) or categories' proportions (for the categorical features). In the description of the weakly supervised target, defined using the presented methodology based on tree-based representations of credal sets, the label y is used as a universally quantified placeholder for any, not explicitly listed, class label.

Feature	Mean	St.Dev.
Age	49.26	16.21
Density (HU)	1246.01	486.43
Volume (mm3)	130,249.46	234,328.63
Diameter (mm)	10.71	5.96
Distance from Mandibular Symphysis (mm)	41.66	15.96
Distance from Mandibular Body (mm)	8.66	4.10
Depth (mm)	5.34	9.86
Side	Left: 48.48 %, Right: 51.52 %	
Shape	Elongated: 59.09 %, Spherical: 40.91 %	
Failure (conversion to sialoadenectomy)	Yes: 3.79 %, No: 96.21 %	
Complications	Yes: 5.30 %, No: 94.70 %	
Target	Transoral: 87.12 %, Endoral: 12.88 %, Sialoadenectomy: 0.0 %	
Weakly Supervised Target	Transoral > y: 79.55 %, Endoral > y: 11.36 %, Sialoadenectomy > Transoral > Endoral: 3.79 %, "?": 5.30 %	

original labels, while the WSL ones were trained using the credal sets obtained as described above. In detail, we applied a 5-fold stratified cross-validation with the constraint that, for each iteration of the cross-validation procedure, the testing fold consisted only of cases associated with a certainly optimal treatment option.[4] We evaluated the models' performance in terms of accuracy, (average) recall and precision, (weighted average) AUC, Brier score and ECI. Differences between the WSL models and the corresponding fully supervised ones were evaluated by means of interval analysis, comparing the 95 % confidence intervals for the reported performance measures. We declared two methods to be significantly different if the corresponding intervals did not overlap. Intervals were computed based on the cross-validation variance and adjusted for multiple comparisons using Bonferroni's method.

Second, we compared the WSL models against a selection of other state-of-the-art benchmark approaches for weakly supervised learning and treatment selection, with the aim of evaluating the potential advantages and limitations of the proposed approach in the context of the specific real-world medical problem considered. We examined three main categories of benchmark methods:

- **Semi-supervised learning models:** These models assume that the ground truth is either certainly known (when no complication or failure occurs), or completely unknown. This corresponds to an extreme setting in which the weakly supervised labels are either of the form T (where Y is the original treatment indication) for cases with no complications or failures, or "?" for cases with complications or failures. We considered two state-of-the-art semi-supervised learning methods: Label Propagation (LP) [36] and Self-Supervised Learning (SSL)[33], using the implementations provided in the `scikit-learn` library;
- **Weakly supervised learning models:** These models use the same representation of weak labels in terms of the possibility distribution $poss_T$ described in Section 2.3, but are based on the Generalized Risk Minimization (GRM) algorithm [9] rather than the RRL algorithm. We used the GRM implementation available in the `scikit-weak` library;
- **Causal inference-based methods:** These models reconstruct the causal structure of the problem—i.e., the probabilistic relationships among features and their influence on treatment outcomes—as a basis for inferring the optimal treatment. We considered two such models: one based on structural learning of Causal Bayesian Networks (BCN) [52], and the generative adversarial network-based GANITE algorithm [56].[5]

As in the previous evaluation, WSL models and benchmarks were compared using 5-fold stratified cross-validation, with the constraint that the test fold in each iteration included only cases associated with a certainly optimal treatment. Model performance was assessed using accuracy, (average) recall and precision, (weighted average) AUC, Brier score, and ECI. Differences were assessed via interval analysis, by comparing the 95 % confidence intervals of the respective performance metrics. Two models were considered significantly different if their confidence intervals did not overlap. Intervals were computed based on the cross-validation variance and adjusted for multiple comparisons using Bonferroni's method.

Finally, we evaluated the generalization performance of the weakly supervised models through a stratified hold-out validation, so as to identify the best base class for RRL, and determine which features could be considered most important for treatment selection. We randomly split the dataset into a training set (80 %) and a test set (20 %), with the constraint that the test set should consist only of cases associated with a certainly optimal treatment. We used the training set to train the models and the test set for validation. Model performance was evaluated in terms of sensitivity, specificity, PPV, NPV, AUC and Brier score. For LR, SVM and DT, to ensure that the learned models were interpretable, we applied a post-processing step on the result of RRL: namely, we used the predicted labels [33] produced by RRL on the training set (that is, to each instance we associated the treatment predicted by RRL) to obtain a new ground truth, which we then used to re-train the interpretable models. We then evaluated these re-trained models on the hold-out test set.

3. Results

The features' distribution for the sample of patients considered in the study is shown in Table 1. Correlations among features are reported in Table 2: all correlations were weak or negligible, except those between Depth and Distance from the Mandibular Symphysis, and Distance from the Mandibular Symphysis and Distance from the Mandibular Body, which were both moderate. The target distribution was strongly imbalanced: out of the total of 132 patients, 105 underwent a transoral asportation, of which 5 had a failure with re-conversion to sialoadenectomy, and 17 underwent an endoral asportation. 7 patients had a complication post-surgery: 2 of these previously underwent an endoral asportation, while 5 underwent a transoral asportation. After transforming the labels into the corresponding weakly supervised targets, the distribution was less markedly imbalanced: for this reason, and to avoid bias in the analysis [15], we decided not to apply any pre-processing technique for dealing with label imbalance. Instead, based also on the

[4] These cases were identified as cases for which there were no complications or failures, and for which the clinical characteristics allowed the clinicians to reliably identify the most appropriate treatment option, which was always a minimally invasive one (either transoral or endoral surgery). While this does not guarantee that the performed treatment was objectively optimal, we argue that it is the strongest admissible proxy for optimality available in retrospective observational data, short of a randomized trial or fully-specified causal inference model (both of which face practical barriers in this domain).

[5] An extended comparison with another class of causal inference methods, namely conditional average treatment effect estimators, is reported in Appendix C.

Table 2
Correlations and associated adjusted p-values (in parentheses) for the features considered in the analysis: DMS denotes distance from mandibular symphysis, DMB denotes distance from mandibular body. Correlations between numerical features were computed using Pearson r coefficient, correlations between numerical and categorical features were computed using point biserial correlation (associated p-values with Mann-Whitney U test), and correlations between categorical features were computed using Cramer's V (associated p-values with the χ^2 test). Significant results are denoted in bold: p-values were adjusted for multiple comparisons using Bonferroni's method.

	Density	Volume	Diameter	DMS	DMB	Depth	Side	Shape
Age	0.024 (1.0)	0.122 (0.733)	0.165 (0.267)	0.017 (1.0)	−0.071 (1.0)	−0.165 (0.266)	−0.035 (1.0)	0.083 (0.754)
Density		0.140 (0.493)	0.151 (0.377)	0.102 (1.0)	0.026 (1.0)	−0.117 (0.822)	0.009 (1.0)	−0.153 (0.588)
Volume			0.220 (0.051)	0.101 (1.0)	−0.137 (0.529)	−0.116 (0.829)	−0.047 (1.0)	−0.065 (1.0)
Diameter				0.193 (0.12)	−0.078 (1.0)	−0.144 (0.446)	−0.058 (1.0)	**0.360 (< 0.001)**
DMS					**0.468 (< 0.001)**	**−0.554 (< 0.001)**	−0.023 (1.0)	−0.168 (0.193)
DMB						**−0.317 (< 0.001)**	0.043 (1.0)	−0.145 (0.82)
Depth							0.082 (1.0)	0.073 (1.0)
Side								0.025 (1.0)
Shape								

current state-of-the-art knowledge on label imbalance in WSL problems [55], we only employed class re-weighting in the training of ML models.

The results of the cross-validation analysis are reported in Fig. 1 and Table 3. In almost all cases, with the exception of the Brier score for RF and SVM, the weakly supervised models reported better cross-validation performance than the corresponding fully supervised learning ones: in almost all cases, this difference was statistically significant.

The results of the comparison with benchmark methods are reported in Fig. 2. In almost all cases—except for the Brier score—the best-performing models were WSL models, although in some instances the differences were not statistically significant. In particular, GANITE did not perform significantly worse than the best WSL models in terms of accuracy, recall, precision, AUC, and Brier score. Semi-supervised learning methods, as well as GRM, were significantly outperformed by the WSL models on most evaluation metrics—with the exception of Brier score and, in the case of semi-supervised learning, precision. Although causal inference-based methods—particularly GANITE—achieved comparable mean performance to the WSL models, they exhibited substantially higher variance, leading to overly wide confidence intervals.

The results of the generalization performance analysis are reported in Fig. 3, in terms of ROC and calibration curves, and Table 4.

The best interpretable models (LR and DT) are illustrated in Figs. 4 and 5, while an explanation of the feature influence for the black-box models (RF and XGB), in terms of SHAP values, is in Fig. 6. In all cases, "distance from the mandibular symphysis" was identified as the most important feature, followed by volume, density, depth, age and "distance from the mandibular body".

Table 3
Results of the cross-validation analysis, along with the corresponding 95 % confidence intervals.

	Accuracy	Recall	Precision	AUC	Brier	ECI
WDT	.78 ±.04	.78 ±.04	.81 ±.03	.66 ±.06	.21 ±.04	.80 ±.02
DT	.68 ±.02	.66 ±.05	.74 ±.04	.61 ±.03	.14 ±.04	.73 ±.01
WLR	.67 ±.04	.67 ±.04	.85 ±.02	.75 ±.03	.26 ±.03	.71 ±.02
LR	.59 ±.05	.55 ±.05	.76 ±.02	.62 ±.05	.28 ±.03	.59 ±.02
WRF	.86 ±.04	.86 ±.04	.83 ±.03	.80 ±.06	.13 ±.02	.77 ±.02
RF	.73 ±.03	.79 ±.05	.74 ±.06	.72 ±.06	.11 ±.03	.70 ±.03
WSVM	.73 ±.04	.73 ±.04	.86 ±.01	.73 ±.07	.12 ±.02	.87 ±.01
SVM	.59 ±.05	.66 ±.04	.79 ±.03	.61 ±.07	.17 ±.03	.76 ±.03
WXGB	.84 ±.03	.84 ±.03	.82 ±.03	.74 ±.04	.14 ±.03	.89 ±.02
XGB	.72 ±.02	.71 ±.03	.68 ±.02	.68 ±.04	.19 ±.02	.76 ±.02

4. Discussion

In this article, we studied the application of ML techniques to the problem of treatment selection. This problem, despite its importance in determining patients' outcomes, has received far less attention than either diagnosis or prognosis, probably due to its inherent and intrinsic complexity [7]. Indeed, while ML methods have been proposed to solve this problem from a methodological point of view, their adoption in practical contexts has so far lagged behind.

To address this gap, in this work we propose to formalize the treatment selection problem as a WSL task. After motivating this position, we describe a lightweight framework by which any treatment selection task can be equivalently described as a credal learning problem. Thus, we provide an algorithmic approach to solve this learning problem and prove that, under weak assumptions, this approach can indeed recover the optimal treatment. Finally, we prove the effectiveness of our proposal in a real-world medical problem, i.e., treatment recommendation for sialolithiasis, showing promising results. In particular, we prove three main results.

First, the proposed WSL approach provides better performance and increased robustness compared to traditional ML approaches: this result has been proven both theoretically, through Theorem 4, as well as empirically on the considered use-case.

Second, the proposed WSL approach can also provide better performance and increased stability in comparison with other state-of-the-art approaches for weakly supervised learning and treatment selection: this result is particularly interesting because it shows that the proposed approach is better able to use the latent information about the optimal treatment compared to other state-of-the-art approaches. On the one hand, our method provides significantly better performance than semi-supervised learning methods: this result is not unexpected, as the proposed approach relies on credal sets to model weak supervision about the optimal treatment in a more flexible way than simply assuming no information is available about the optimal treatment (as in semi-supervised learning). On the other hand, and more interestingly, the proposed approach also out-performed other weakly supervised learning methods (i.e., GRM), as well as causal inference methods. In regard to the comparison with GRM, we observe that already in more restricted forms of weakly supervised learning (namely, learning from fuzzy labels), a variant of RRL has been shown to outperform GRM [9], primarily due to the fact that solving the optimization problem underlying GRM is computationally hard, and existing approximation methods trade-off accuracy with reduced complexity. A similar explanation can be given to motivate the observed difference in performance between the proposed WSL methods and causal inference approaches, which furthermore rely on strong assumptions that, when not satisfied, can lead to sub-optimal performance.

Finally, the proposed approach can obtain satisfactory generalization power as well as provide relevant clinical insights about sialoliths' morphological features that can be used to effectively guide treatment, thus providing positive answers to both our research questions.

Focusing more in detail on the considered case study: over the years, various decision-making algorithms have been proposed based on single-center experiences [5,28,29,40]. These algorithms were obtained

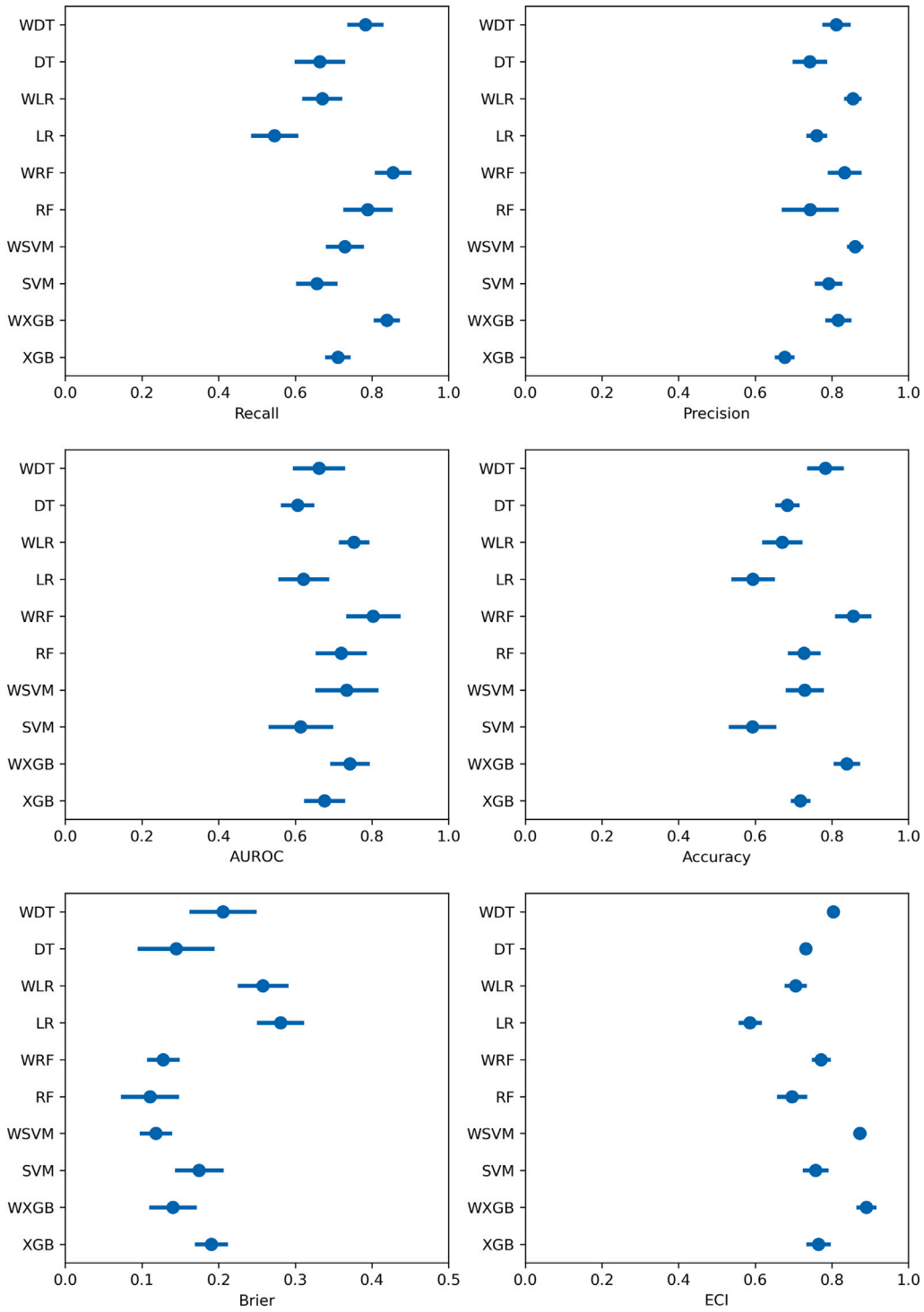

Fig. 1. Results of the cross-validation analysis. Significant differences among algorithms are defined in terms of interval analysis: two algorithms are deemed to have significantly different performance if the corresponding 95 % confidence intervals do not overlap.

through traditional, knowledge-driven approaches rather than ML techniques, and led to suboptimal results and poor generalizability in subsequent studies [12]. Thus, the aim of our case study was to predict the optimal treatment for patients with submandibular stones from radiological information easily retrieved from CT scans, using modern ML approaches. Target information was therefore considered to be the type of surgical procedure employed and any complications or conversions from transoral conservative surgery to sialoadenectomy that had

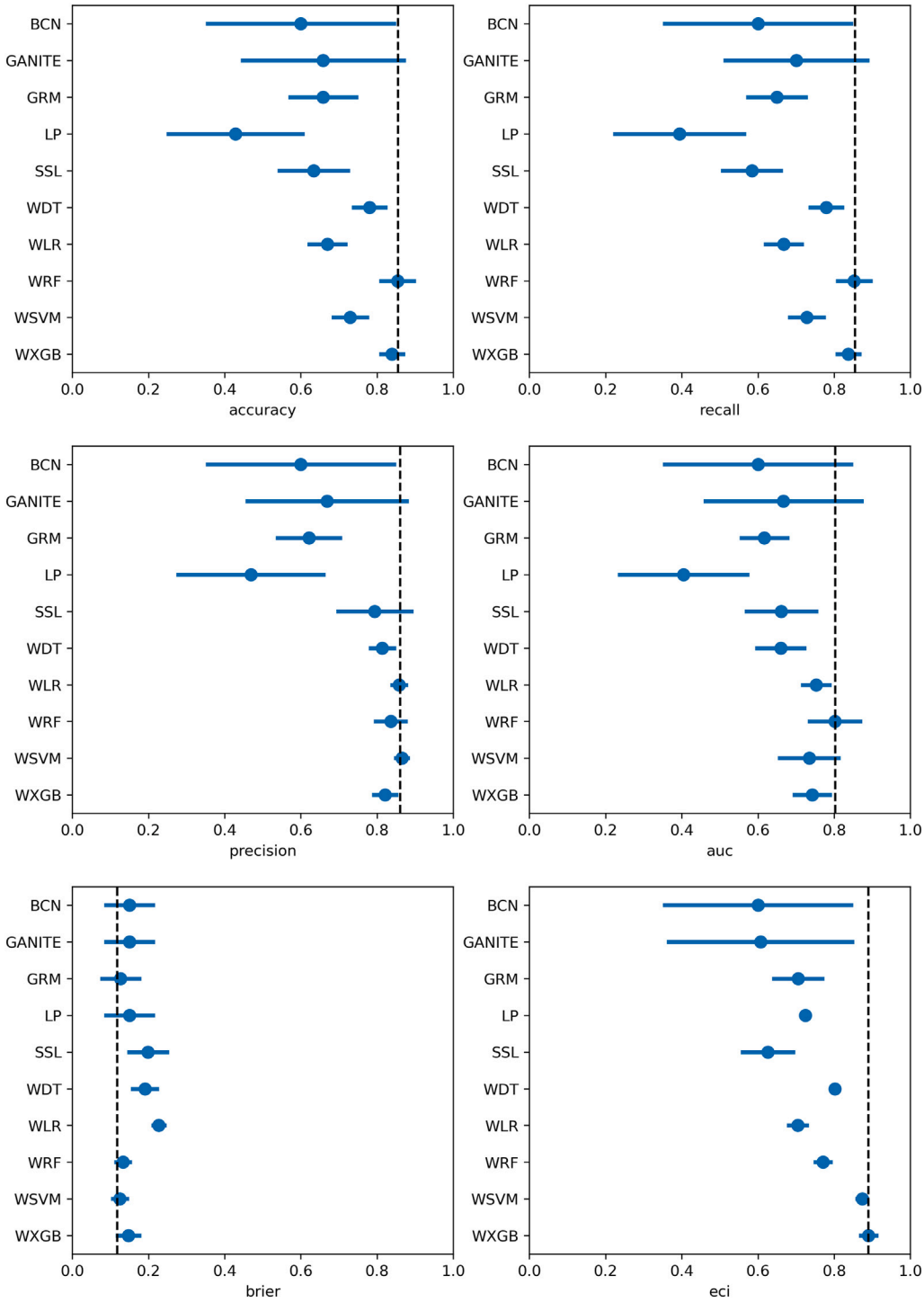

Fig. 2. Results of the benchmark comparison analysis. Significant differences between algorithms are determined via interval analysis: two algorithms are considered to have significantly different performance if their 95 % confidence intervals do not overlap. In each plot, the dashed line indicates the performance of the best WSL model.

taken place, which have all been employed and combined to define a weakly supervised target.

The results of our case study show how, even though there are no significant differences between the WSL models, all of them outperformed the corresponding fully supervised models. Overall, the best performing models, particularly in terms of area under the ROC curve and sensitivity, were RRL based on the black-box RF and XGB models. Fig. 3 provides a general overview of the accuracy of each model in predicting

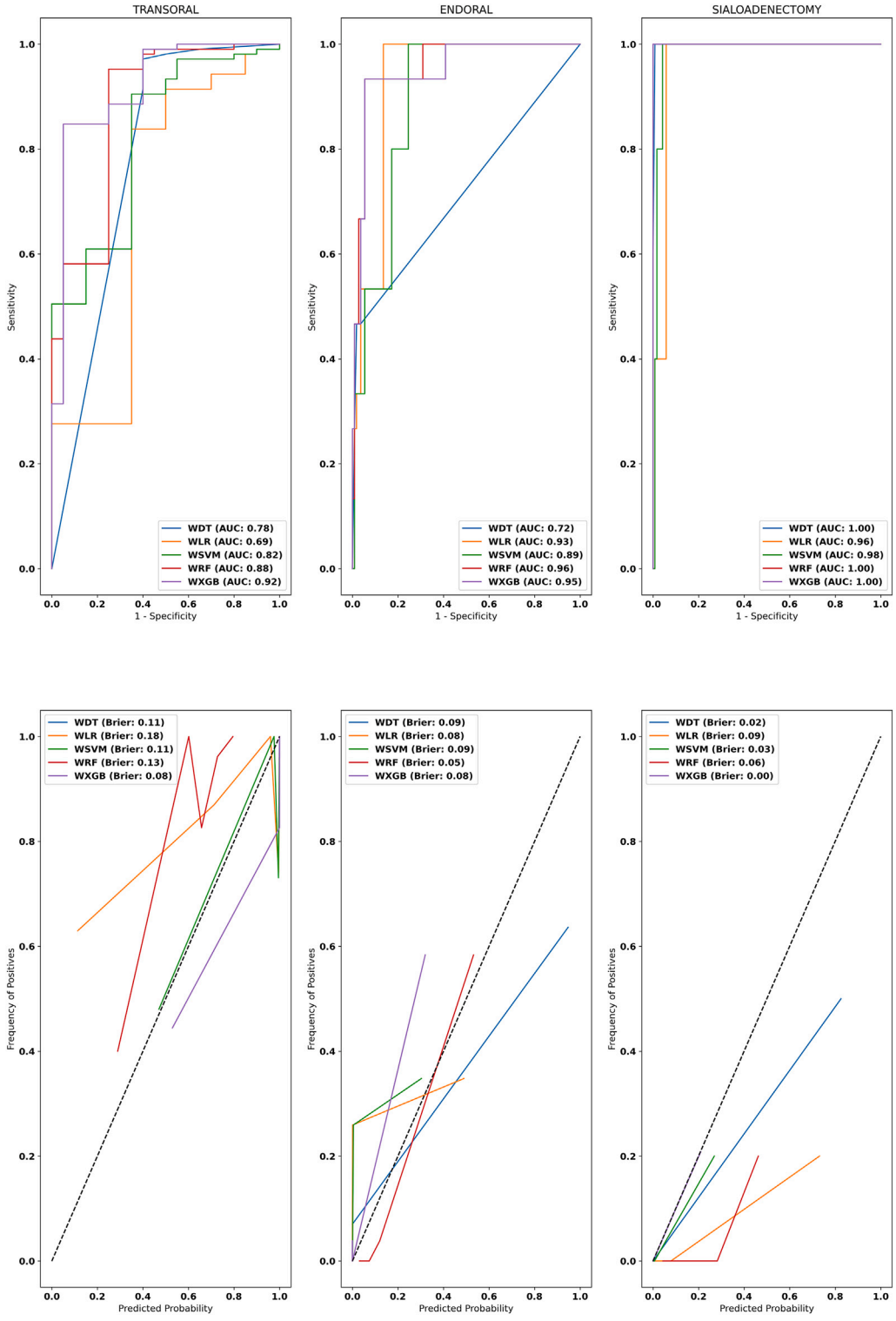

Fig. 3. Results of the generalization analysis for the weakly supervised learning algorithms, represented in terms of class-wise ROC curves (top) and calibration curves (bottom).

each surgical option. Despite the presence of a large imbalance among the three classes due to the limited number of sialoadenectomy cases, for this latter class, both DT, XGB and RF models exhibit high sensitivity and specificity, albeit at the cost of a lower positive predictive value, which was nonetheless (and especially so for XGB) sufficiently high. In this regard, current sialoadenectomy rates described in literature range from 5 to 10 % [13,28], while in the present patient cohort this procedure was less frequently employed: this could be due to the fact that patients were operated on by a surgeon with extensive experience in the field of obstructive salivary disease. Notably, current

Table 4
Results of the generalization analysis for the weakly supervised learning algorithms, along with the corresponding 95 % confidence intervals.

Class	Model	Sensitivity	Specificity	PPV	NPV	AUC	Brier
Transoral	WDT	0.95 ± 0.01	0.6 ± 0.05	0.93 ± 0.01	0.71 ± 0.04	0.78 ± 0.03	0.1 ± 0.02
	WLR	0.82 ± 0.03	0.65 ± 0.04	0.92 ± 0.01	0.41 ± 0.05	0.68 ± 0.04	0.17 ± 0.03
	WSVM	0.94 ± 0.01	0.5 ± 0.05	0.91 ± 0.02	0.62 ± 0.04	0.82 ± 0.03	0.11 ± 0.02
	WRF	0.93 ± 0.01	0.75 ± 0.04	0.95 ± 0.01	0.68 ± 0.04	0.87 ± 0.02	0.13 ± 0.02
	WXGB	0.98 ± 0.0	0.6 ± 0.05	0.93 ± 0.01	0.86 ± 0.02	0.93 ± 0.01	0.07 ± 0.01
Endoral	WDT	0.67 ± 0.11	0.98 ± 0.01	0.64 ± 0.12	0.93 ± 0.03	0.72 ± 0.1	0.09 ± 0.04
	WLR	0.63 ± 0.12	0.96 ± 0.02	0.67 ± 0.11	0.94 ± 0.03	0.92 ± 0.04	0.08 ± 0.04
	WSVM	0.43 ± 0.12	0.95 ± 0.02	0.71 ± 0.1	0.92 ± 0.04	0.89 ± 0.05	0.09 ± 0.04
	WRF	0.77 ± 0.09	0.96 ± 0.02	0.67 ± 0.11	0.95 ± 0.02	0.96 ± 0.02	0.05 ± 0.03
	WXGB	0.67 ± 0.11	0.99 ± 0.01	0.78 ± 0.09	0.93 ± 0.03	0.97 ± 0.02	0.07 ± 0.03
Sialoadenectomy	WDT	1.0 ± 0.0	0.99 ± 0.01	0.83 ± 0.12	1.0 ± 0.0	1.0 ± 0.0	0.01 ± 0.01
	WLR	1.0 ± 0.0	0.88 ± 0.1	0.25 ± 0.16	1.0 ± 0.0	0.96 ± 0.03	0.08 ± 0.07
	WSVM	1.0 ± 0.0	0.97 ± 0.03	0.56 ± 0.22	1.0 ± 0.0	0.98 ± 0.02	0.03 ± 0.02
	WRF	1.0 ± 0.0	0.98 ± 0.01	0.71 ± 0.18	1.0 ± 0.0	1.0 ± 0.0	0.06 ± 0.05
	WXGB	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	0.0 ± 0.0

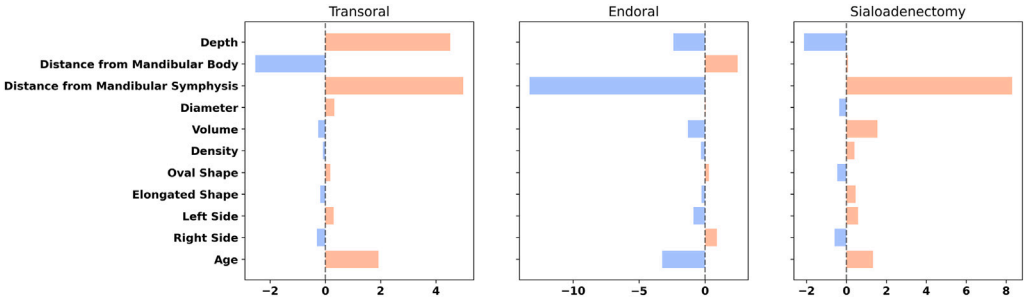

Fig. 4. Graphical representation of the feature importance for the different features for the LR model, stratified by class. For each class and feature, the color denotes the sign of the corresponding coefficient and hence its effect on the probability associated to the corresponding class: thus, red denotes an increase in the probability associated to that class, while blue denotes a reduction in probability. The x-axis is depicted in the scale of relative log-odds.

parameters used for surgical planning of submandibular stones are the stone's diameter and clinical parameters such as stone palpability and endoscopic accessibility [5,28,29]. The discussed WSL ML models have instead used new parameters, as shown in Figs. 4–6, such as the stone's volume and density, which reflect its composition, or the stone's distance from the mandible, introducing new cut-offs and thresholds that might help physicians in the surgical planning of this condition and potentially lead to the definition of new guidelines and decision algorithms. Current clinical guidelines for the treatment of submandibular sialolithiasis do not distinguish between simple and complex cases - such as those involving friable, crumbly stones that cannot be removed intact and may leave residual stone fragments within the ductal tree. These remnants can later become symptomatic, leading patients to seek additional medical intervention. The new parameters incorporated by our proposed machine learning models—such as stone density and volume—may help address this limitation by supporting more precise and comprehensive surgical planning.

Strengths of the present study include: the development of a novel methodological approach that, other than showing promising results from an empirical point of view, provides theoretical guarantees on the recovery of the optimal treatment recommendation; the usage of a highly curated ground truth dataset; as well as providing interpretable insights on which features influence the selection of the most appropriate treatment, also relying on novel parameters such as sialoliths' volume and density. We believe that all of these insights can be particularly useful for designing further studies that on the one hand aim at applying a WSL paradigm to solve treatment selections tasks in other settings, on the other hand aim at validating new procedures and guidelines for surgical planning in the context of sialolithiasis inspired by the above discussed findings.

The main limitation of the present study, by contrast, concerns the limited size of the considered dataset as well as the corresponding label imbalance. In regard to this latter, however, we note that the developed weakly supervised models (and especially random forest, extreme gradient boosting and decision tree) were able to provide satisfactory sensitivity, specificity and PPV both for the more common as well as the rarer treatment options, thus showing that the proposed WSL approach can also be useful to handle imbalanced learning problems in a robust manner. In regard to the limited dataset size, while we note that such a limitation is common when one considers less common pathological conditions that go beyond the traditional settings studied in ML contexts [2] (e.g., oncology, radiology or diabetes), as well as the fact that even when considering the lower extremum of the confidence intervals for performance reported in Table 4 the developed WSL models mostly showed acceptable performance, we believe that further studies should validate the generalizability of our clinical findings. This applies, in particular, to out-of-domain evaluations that assess the generalizability of our results across centers, operators and imaging devices, which are of fundamental importance for clinical validity. Finally, the primary focus of this study was on cone beam computed tomography, yet future studies could also incorporate ultrasonography data into machine learning models, providing a more comprehensive tool for treatment planning by integrating detailed gland and ductal information.

More generally, although this study focused on the application of the proposed methodology to the treatment selection of submandibular stones, future research should investigate its applicability in other clinical settings and domains to assess its generalizability. It is important to note that the proposed method—particularly the credal encoding of epistemic states - requires the explicit annotation of treatment failures or conversions. In the absence of such information, applying the method

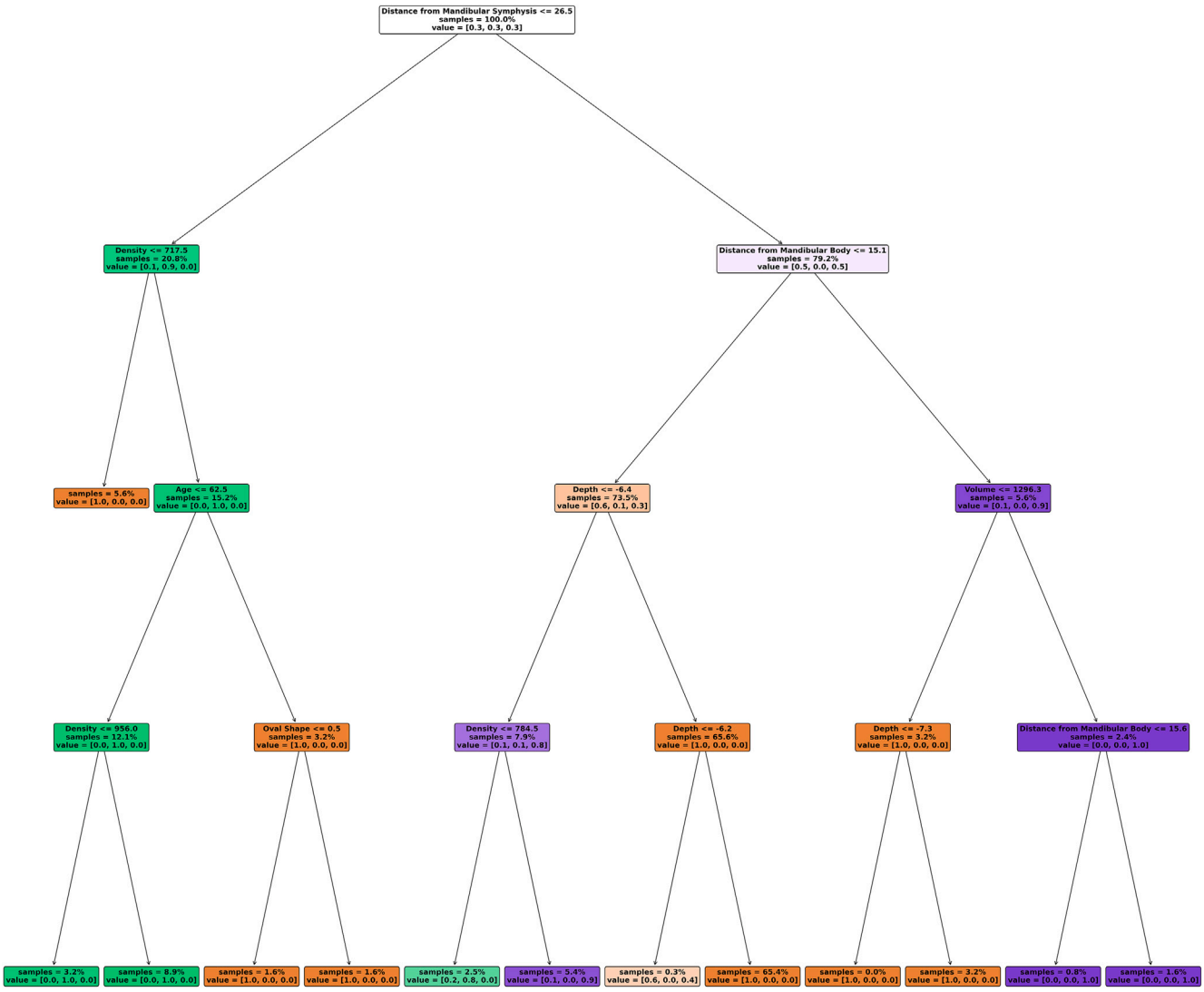

Fig. 5. Graphical representation of the decision tree model. For each node, we visualize: the split feature and corresponding feature value, the percentage of samples corresponding to the path connecting the root to the node, and the proportion of instances belonging to each class in that node. For each node, the color represents the class with highest associated proportion: orange corresponds to transoral, green to endoral and blue to scialoadenectomy.

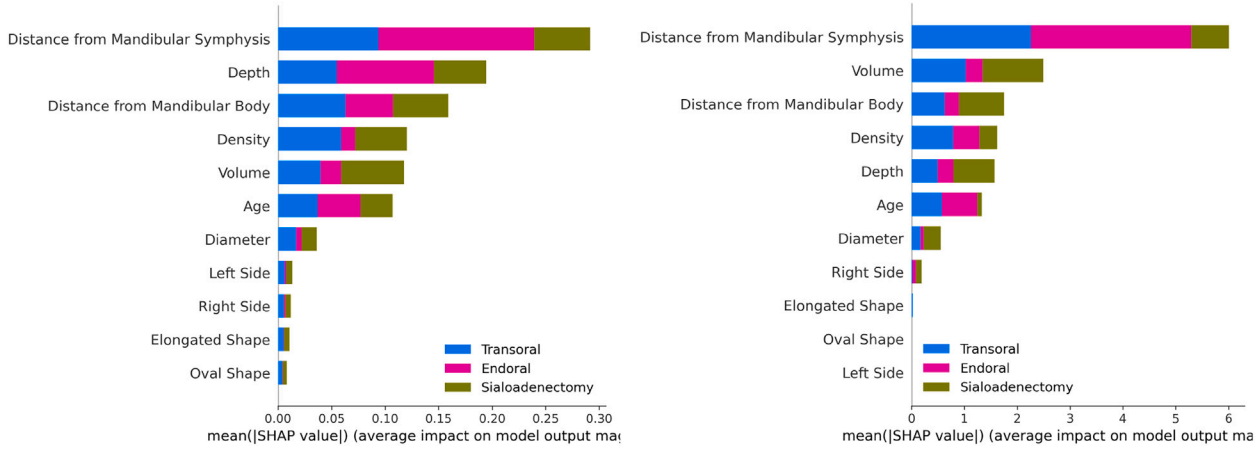

Fig. 6. Feature importance for the random forest (left) and XGBoost (right) models, computed in terms of SHAP values and represented in terms of stacked bar charts: for each feature, the stacked bar charts denote the contribution of that feature to the probability scores of the different classes.

becomes difficult or even infeasible. In this regard, our positive results underscore the importance of systematically collecting and sharing this type of data, as it enables the deployment of advanced strategies for treatment selection.

5. Conclusions

To the best of our knowledge, the present study represents the first application of WSL to address the treatment selection problem, as well as the first study to apply machine learning models in the treatment of sialolithiasis. After developing a lightweight but theoretically justified framework, based on a credal learning approach, to formalize treatment selection as a WSL problem, and having provided conditions under which such tasks could be solved along with precise generalization guarantees, we have applied our methodology in the setting of sialolithiasis. Five WSL ML models were developed, including both black-box and interpretable models. In all cases, the WSL models outperformed the corresponding traditional fully supervised approaches, as well as state-of-the-art weakly supervised and treatment selection methods, showing how our proposed methodology can improve performance and robustness. Furthermore, the black-box Random Forest and eXtreme Gradient Boosting models achieved the best results in terms of area under the ROC curve and sensitivity. These models introduced new parameters and thresholds for surgical planning, which could be employed by clinicians to more appropriately address treatment selection. While our empirical analysis is intended as a proof of concept rather than a definitive assessment of the method's practical advantages, our results demonstrate the feasibility of using a WSL approach for treatment-selection problems and lay the groundwork for more extensive evaluations—across additional simulations and diverse real-world datasets—and for applications in other clinical settings, ultimately providing more comprehensive validation in practice.

CRediT authorship contribution statement

Andrea Campagner: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Matteo Lazzeroni:** Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Data curation, Conceptualization. **Caterina Pizzi:** Writing – review & editing, Validation, Data curation. **Caterina Sattin:** Writing – review & editing, Validation, Data curation. **Giulia Buccichini:** Writing – review & editing, Validation, Data curation. **Massimo Del Fabbro:** Writing – review & editing, Visualization, Supervision, Methodology. **Gianluca Martino Tartaglia:** Writing – review & editing, Visualization, Supervision, Methodology. **Maria Cristina Firetto:** Writing – review & editing, Methodology, Data curation. **Gianpaolo Carrafiello:** Writing – review & editing, Validation, Supervision. **Michael Koch:** Writing – review & editing, Validation. **Pasquale Capaccio:** Writing – review & editing, Supervision, Project administration, Conceptualization. **Federico Cabitza:** Writing – review & editing, Visualization, Validation, Supervision, Project administration, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Additional results and proofs

Theorem 1. *Let Y be a finite set. Let T be a tree whose nodes are labeled with elements from some subset $Y' \subseteq Y$. We say that a probability distribution P is compatible with T if $(y, y') \in T \implies P(y) \geq P(y')$. Then, the set $Q(T)$ of probability distributions compatible with P is a credal set (that is, it is convex and closed). Furthermore, $Q(T)$ is the set of probability distributions such that:*

1. *$P(y) \leq 1$ if y is the root of T or y is not in T;*
2. *$P(y') \leq \frac{1}{d}$, where d is the depth of y' in T, otherwise.*

Thus, each tree T can be associated with a normalized possibility distribution [20] $poss_T : Y \to [0, 1]$ defined by $poss_T(y) = max_{P \in Q(T)} P(y)$.

Proof. The fact that the set of probability distributions compatible with T is a credal set derives from Miranda and Destercke [43]. For the second part, let $y \in Y$ be the root of T: clearly, the distribution s.t. $P(y) = 1$ and for all other $y' \in Y$, $P(y') = 0$ belongs to $Q(T)$. The same holds if y does not belong to T (as, in this case, we have no constraint at all on the value of $P(y)$). Thus, let y be an internal node of T, and let d be the depth of y. Clearly, if $P \in Q(T)$, then if y' is an ancestor of y in T it cannot hold that $P(y) > P(y')$. However, the distribution P s.t. $P(y) = \frac{1}{d}$ and, for all ancestors y' of y, $P(y') = \frac{1}{d}$ clearly is in $Q(T)$. □

Corollary 1. *Let T be such that either:*

1. *T is empty;*
2. *There exist $y, y' \in Y$ s.t. T contains only the arcs (y', y) and for all $y'' \neq y' \in Y$, (y, y'');*
3. *There exists $y \in Y$ s.t. T contains all and only the arcs (y, y') for all $y' \in Y$.*

Then, the possibility distributions constructed from T can be defined as follows:

1. *For each $y \in Y$, $poss_T(y) = 1$;*
2. *$poss_T(y') = 1$, $poss_T(y) = \frac{1}{2}$ and for all other $y \in Y$, $poss_T(y'') = \frac{1}{3}$;*
3. *$poss_T(y) = 1$ and for all other $y' \in Y$, $poss_T(y) = \frac{1}{2}$.*

Proof. The result directly follows from Theorem 1. □

Proposition 1. *Let H be a base hypothesis class. Let A be a polynomial-time learning algorithm for H. Then, Algorithm 1 has time complexity $O(k * T_A(|S|) + |S||Y|)$, where $T_A(|S|)$ is the time complexity of A on a dataset of size $|S|$.*

Proof. The possibility distribution $poss_T$ can be easily constructed in time $O(|Y|)$ for each instance in the training set S, by applying Corollary 1, thus requiring a total time of $O(|S||Y|)$. Similarly, if training a base classifier on a training set S requires time $T_A(|S|)$, then the total time to train the RRL ensemble is $O(k * T_A(|S|))$, from which the result follows. □

Theorem 2. *Assume that, with probability 1 over the sampling of an imprecise instance $(x, y, Q(x))$ drawn from M, where $Q(x)$ is constructed as in Corollary 1, it holds that $P(\cdot|x) \in Q(x)$. Assume, further, that P is realizable for $\mathcal{H}$ (i.e., there exists $h \in \mathcal{H}$ s.t. with probability 1 $h(x) = P(\cdot|x)$). Then, asymptotically (i.e., when the sample size n and ensemble size k go to infinity), it holds that $L_P(RRL) = min_f L_P(f^*)$, where f ranges over the functions measurable with respect to P. That is, asymptotically, RRL converges to a Bayes predictor.*

Proof. The result directly follows from Corollary 3.2 in [9], when applied to the possibility distribution $poss_T$ constructed as in Corollary 1 and using the version of RRL given in Algorithm 1, noting that, due to realizability, the weight of any model h s.t. $h(x) = P(\cdot|x)$ is maximal. □

Appendix B. Robustness to noisy labels

In this section we describe additional experiments, conducted in a simulated setting, aimed at evaluating the robustness of the proposed WSL methodology to deviations from the assumptions in Theorem 2, namely the assumption that $P(\cdot|x)$ (the true optimal treatment distribution) lies in $Q(T)$ (the credal set constructed from the epistemic stance of the clinician).

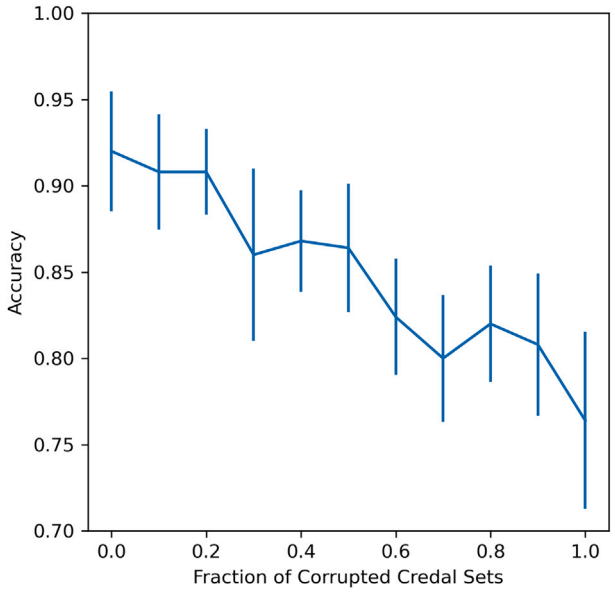

Fig. B.7. Results of the robustness analysis for the RRL algorithm, showing the average performance and 95 % confidence intervals in the simulated credal learning scenario. Confidence intervals were generated by considering the distribution of accuracy scores across the 100 repetitions of the simulation experiment.

To this aim, we devised a simulated experiment in which binary classification data were generated according to a Gaussian mixture model (using a separate Gaussian for each of the two classes, 0 and 1), using the `scikit-learn` `make_blobs` function Specifically, we generated a dataset encompassing 100 instances (50 for each class) and 2 continuous features: this dataset was then split into a training set and test set, according to an 80/20 split. For each instance x in the training set, a credal set satisfying the above mentioned assumption was generated by setting $poss(y) = 1$ if $P(y|x) \geq 0.5$ (where P is the probability distribution defined by the adopted Gaussian mixture model), and otherwise setting $poss(y) = P(y|x)$: it is easy to see that the credal set $Q(T)$ corresponding to the constructed possibility distribution $poss$ always contains $P(y|x)$. Violations of the assumptions were generated by changing k% of the created credal sets (with k ranging in $[0, 10, 20, \dots, 100]$) according to the following procedure: first, one of two classes y was randomly selected with equal probability; second, a number r was selected uniformly at random in $[0, 1]$; finally, $poss(y)$ was set to r and $poss(1 - y)$ was set to 1. Finally, an RRL model (using decision trees with default hyperparameters as base model) was trained on the so-constructed corrupted training set and subsequently evaluated, in terms of accuracy, on the hold-out test set. The above simulation process was repeated 100 times, in order to evaluate the robustness of the RRL algorithm: the results of this analysis are reported in Fig. B.7.

In all cases, while the performance of the RRL algorithm significantly decreased with increasing fractions of corrupted credal sets (hence, larger violations of the above mentioned assumption), nonetheless, the RRL algorithm always reported an accuracy significantly larger (as assessed by interval analysis) than 70 %: this holds also in the extreme case where 100 % (that is, all) credal sets were corrupted according to the above described procedure. Thus, this result illustrates the robustness of the considered RRL algorithm showing its ability to effectively recover the correct class even in the presence of systematic violations of the assumptions in Theorem 2. Nonetheless, as this result applies only to a relatively simple, simulated setting, we leave to future work further exploration of the failure conditions for the proposed approach.

Appendix C. Comparison with CATE inference methods

In this section, we relate our framework to the potential outcomes framework for causal inference and compare the proposed RRL algorithm against state-of-the-art methods for conditional average treatment effect (CATE) estimation.

Following the terminology introduced in Section 2, let X be the feature space, and Y the treatment space. Additionally let Z be the outcome space, that is, the set of possible outcomes resulting from the application of any treatment $y \in Y$: we will assume that $Z = \{0, 1\}$, where $Z = 1$ denotes the absence of any complications or failures. Let $\mathcal{D}$ be a distribution over $X \times Y \times (Z)^{|Y|}$: the quantity $\mathcal{D}(Z^y = 1|x) = \mathbb{E}[Z^y|x]$ (i.e., the probability of the y-th entry of $Z^{|Y|}$ given $x \in X$) is called the *average potential outcome* under treatment y (conditional on x, and given a realization $r = (x, y, (z_1, \dots, z_{|Y|}))$, z_y is called *potential outcome* for individual r. The aim of causal inference methods is to recover the vector $(\mathbb{E}[Z^1|x], \dots, \mathbb{E}[Z^{|Y|}|x])$. A key assumption for this to be possible is exchangeability (i.e., $Y \perp (Z^1, \dots, Z^{|Y|})|X$). Under such assumptions a variety of methods have been developed with the goal of estimating the average potential outcomes[6]: in particular, state-of-the-art approaches are based on the meta-learner principle, in which a base class of machine learning methods is employed to construct a CATE estimator by combining multiple base models [32].

In Section 2, our focus is on treatment selection, that is the identification of the treatment that results in the best outcome. We denote the conditional probability that y is the optimal treatment (conditional on x) as $P(y|x)$. In the language of the potential outcomes framework, the quantity $P(y|x)$ can be expressed as:

$$P(y|x) = \mathcal{D}(\forall y' \in Y, Z^{y'} < Z^y|x) \tag{C.1}$$

$$+ \sum_{k=2}^{|Y|} \frac{1}{k} \mathcal{D}(\exists ! \{y_1, \dots y_{k-1}\} \subseteq Y, Z^{y_1} = \dots Z^{y_{k-1}} = Z^y, Z^y < Z^y). \tag{C.2}$$

Consequently, given a CATE estimator, the optimal treatment distribution can be easily recovered by noting that (C.1) is equal to

$$\mathbb{E}[Z^y|x] \cdot \Pi_{y' \neq y}(1 - \mathbb{E}[Z^{y'}|x]),$$

and each term of (C.2) is equal to

$$\mathbb{E}[Z^y|x] \cdot \Pi_{y' \in \{y_1, \dots, y_{k-1}\}} \mathbb{E}[Z^{y'}|x] \cdot \Pi_{y'' \notin \{y, y_1, \dots, y_{k-1}\}}(1 - \mathbb{E}[Z^{y''}|x]).$$

Thus, while in the proposed credal learning framework the aim is to directly infer the optimal treatment distribution, by using the observed outcome (or, more generally, any information about the preferences among treatments, including the subjective belief of the clinician) to relax the (potentially noisy) distribution $L(\cdot|x)$ into a credal set $Q(x)$, causal inference addresses the same problem in an indirect way, by first estimating the potential outcomes and then using these latter to compute $P(\cdot|x)$. While previous work has shown that any causal inference query (hence, also CATE estimation) can be reduced to a credal inference query, given knowledge of the graphical model underlying $\mathcal{D}$, the relationship between credal learning and CATE estimation has not been studied before: therefore, the relationship between the assumptions needed for the successful application of these frameworks is not known. We aim to investigate this relationship in future work.

Aside from the above formal connections between our proposed framework and causal inference, we compared the proposed RRL method against several state-of-the-art meta-learners for CATE estimation: in particular, we considered the T-learner and S-learner, discussed

[6] Formally, a CATE estimator, is a method that, for every $y, y' \in Y$ and $x \in X$, is able to estimate the quantity $\mathbb{E}[Z^y - Z^{y'}|x]$. Clearly, any method capable of estimating the average potential outcome vector is a CATE estimator.

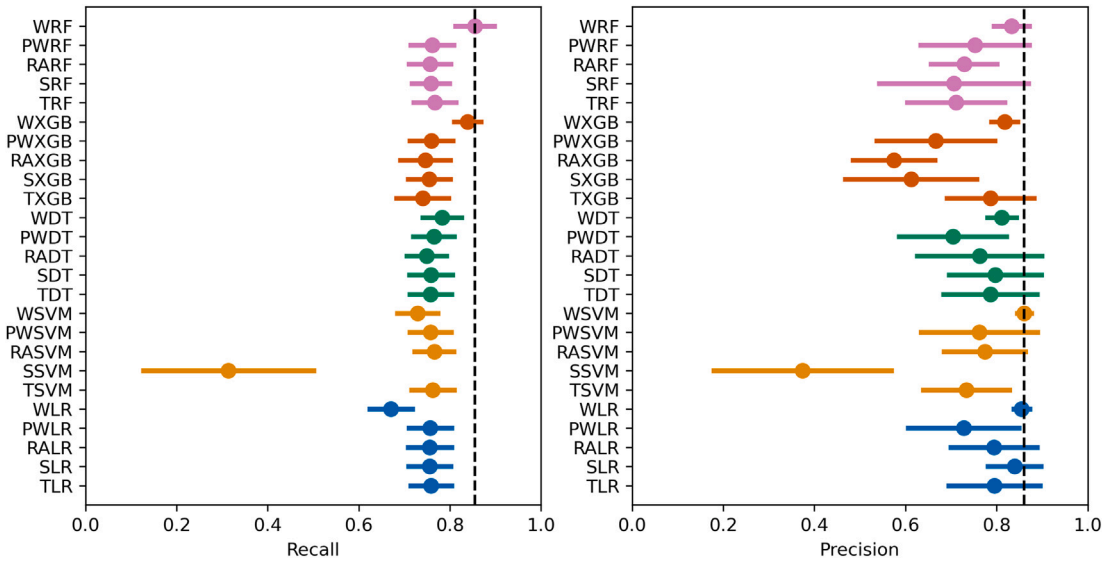

Fig. C.8. Results of the comparison analysis with CATE estimators. Significant differences between algorithms are determined via interval analysis: two algorithms are considered to have significantly different performance if their 95 % confidence intervals do not overlap. In each plot, the dashed line indicates the performance of the best WSL model. Different colors denote different base classes of learner: RF (pink), XGB (red), DT (green), SVM (yellow), LR (blue).

in [32], as well as the RA-learner[7] and PW-learner, discussed in [18]. The comparison between the different methods was performed by adopting the same setting as described in Section 2.4.4: in particular, for every base learner class (DT, SVM, RF, XGB, LR), we compared RRL based on the given class, with the T-learner, S-learner, RA-learner and PW-learner based on the same class (for any model class C, these are denoted respectively as TC, SC, RAC and PWC). Comparisons among methods are performed by means of interval analysis, as described in Section 2.4.4. The results of this analysis are reported in Fig. C.8 in terms of recall and precision. In terms of recall, for most base classes there was no significant difference between RRL and CATE estimators, with the exception of RF and XGB, for which RRL had significantly better performance: in particular, RF-based RRL was the method that reported the highest recall. For non-parametric methods (RF, XGB and DT), RRL reported higher recall than CATE estimators, while for linear methods (LR and SVM) CATE estimators (with the exception of SSVM) outperformed RRL. For precision, while RRL always outperformed the CATE estimators, the difference with the best CATE estimator (PW-learner for RF, T-learner for XGB, S-learner for DT and LR, RA-learner for SVM) was never significant. In general, with the exception of the S-learner based on SVM, we observed no significant differences among CATE estimators, neither for recall nor for precision.

References

[1] A. Ajalloeian, S. Stich, Analysis of SGD with biased gradient estimators, arXiv preprint arXiv:2008.00051, 2020.

[2] A. Althnian, D. AlSaeed, H. Al-Baity, A. Samha, A.B. Dris, N. Alzakari, A. Abou Elwafa, H. Kurdi, Impact of dataset size on classification performance: an empirical evaluation in the medical domain, Appl. Sci. 11 (2021) 796.

[3] D. Ashby, A.F. Smith, Evidence-based medicine as Bayesian decision-making, Stat. Med. 19 (2000) 3291–3305.

[4] T. Augustin, F.P. Coolen, G. De Cooman, M.C. Troffaes, Introduction to Imprecise Probabilities, vol. 591, John Wiley & Sons, 2014.

[5] I. Badash, J. Raskin, M. Pei, L. Soldatova, C. Rassekh, Contemporary review of submandibular gland sialolithiasis and surgical management options, Cureus 14 (2022).

[7] We note that the RA-learner proposed in [18] is a special case of X-learner proposed in [32]: as the former is simpler to implement, as it does not require specifying a weighting function, and, under appropriate assumptions, is similarly asymptotically valid, we do not consider general X-learners in our comparison.

[6] R.A. Berk, L. Brown, E. George, E. Pitkin, M. Traskin, K. Zhang, L. Zhao, What you can learn from wrong causal models, Handb. Causal Anal. Soc. Res. (2013) 403–424.

[7] I. Bica, A.M. Alaa, C. Lambert, M. Van Der Schaar, From real-world patient data to individualized treatment effects using machine learning: current and future methods to address underlying challenges, Clin. Pharmacol. Ther. 109 (2021) 87–100.

[8] A. Campagner, Learnability in “learning from fuzzy labels”, in: 2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IEEE, 2021, pp. 1–6.

[9] A. Campagner, Learning from fuzzy labels: theoretical issues and algorithmic solutions, Int. J. Approx. Reason. (2023) 108969.

[10] A. Campagner, L. Famiglini, A. Carobene, F. Cabitza, Everything is varied: the surprising impact of instantial variation on ML reliability, Appl. Soft Comput. 146 (2023a) 110644.

[11] A. Campagner, et al., Credal learning: weakly supervised learning from credal sets, Front. Artif. Intell. Appl. 372 (2023b) 327–334.

[12] P. Capaccio, F. Ottaviani, R. Manzo, A. Schindler, B. Cesana, Extracorporeal lithotripsy for salivary calculi: a long-term clinical experience, Laryngoscope 114 (2004) 1069–1073.

[13] P. Capaccio, S. Torretta, L. Pignataro, The role of adenectomy for salivary gland obstructions in the era of sialendoscopy and lithotripsy, Otolaryngol. Clin. North Am. 42 (2009) 1161–1171.

[14] H. Chen, A. Shah, J. Wang, R. Tao, Y. Wang, X. Xie, M. Sugiyama, R. Singh, B. Raj, Imprecise label learning: a unified framework for learning with various imprecise label configurations, arXiv preprint arXiv:2305.12715, 2023.

[15] W. Chen, K. Yang, Z. Yu, Y. Shi, C. Chen, A survey on imbalanced learning: latest research, applications and future directions, Artif. Intell. Rev. 57 (2024) 1–51.

[16] T.H. Cormen, C.E. Leiserson, R.L. Rivest, C. Stein, Introduction to Algorithms, MIT Press, 2022.

[17] A. Curth, R.W. Peck, E. McKinney, J. Weatherall, M. van Der Schaar, Using machine learning to individualize treatment effect estimation: challenges and opportunities, Clin. Pharmacol. Ther. 115 (2024) 710–719.

[18] A. Curth, M. Van der Schaar, Nonparametric estimation of heterogeneous treatment effects: from theory to learning algorithms, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 1810–1818.

[19] A.P. Dawid, Causal inference without counterfactuals, J. Am. Stat. Assoc. 95 (2000) 407–424.

[20] D. Dubois, H. Prade, Possibility theory, in: Granular, Fuzzy, and Soft Computing, Springer, 2023, pp. 859–876.

[21] M. Galdermans, B. Gemels, Success rate and complications of sialendoscopy and sialolithotripsy in patients with parotid sialolithiasis: a systematic review, Oral Maxillofac. Surg. 24 (2020) 145–150.

[22] M. Goncalves, M. Schapher, H. Iro, W. Wuest, K. Mantsopoulos, M. Koch, Value of sonography in the diagnosis of sialolithiasis: comparison with the reference standard of direct stone identification, J. Ultrasound Med. 36 (2017) 2227–2235.

[23] L. Grinsztajn, E. Oyallon, G. Varoquaux, Why do tree-based models still outperform deep learning on typical tabular data? Adv. Neural Inf. Process. Syst. 35 (2022) 507–520.

[24] S.R. Herrle, E.C. Corbett Jr, M.J. Fagan, C.G. Moore, D.M. Elnicki, Bayes’ theorem and the physical examination: probability assessment and diagnostic decision making, Acad. Med. 86 (2011) 618–627.

[25] E. Hüllermeier, Learning from imprecise and fuzzy observations: data disambiguation through generalized loss minimization, Int. J. Approx. Reason. 55 (2014) 1519–1534.
[26] K.F. Hung, Q.Y.H. Ai, L.M. Wong, A.W.K. Yeung, D.T.S. Li, Y.Y. Leung, Current applications of deep learning and radiomics on CT and CBCT for maxillofacial diseases, Diagnostics 13 (2022) 110.
[27] R.C. Kessler, R.M. Bossarte, A. Luedtke, A.M. Zaslavsky, J.R. Zubizarreta, Machine learning methods for developing precision treatment rules with observational data, Behav. Res. Ther. 120 (2019) 103412.
[28] M. Koch, K. Mantsopoulos, S. Müller, M. Sievert, H. Iro, Treatment of sialolithiasis: what has changed? an update of the treatment algorithms and a review of the literature, J. Clin. Med. 11 (2021) 231.
[29] M. Koch, J. Zenk, H. Iro, Algorithms for treatment of salivary gland obstructions, Otolaryngol. Clin. North Am. 42 (2009) 1173–1192.
[30] S. Kraaij, K. Karagozoglu, T. Forouzanfar, E. Veerman, H. Brand, Salivary stones: symptoms, aetiology, biochemical composition and treatment, Br. Dent. J. 217 (2014) E23–E23.
[31] K. Kulasekera, S. Tholkage, M. Kong, Personalized treatment selection using observational data, J. Appl. Stat. 50 (2023) 1115–1127.
[32] S.R. Künzel, J.S. Sekhon, P.J. Bickel, B. Yu, Metalearners for estimating heterogeneous treatment effects using machine learning, in: Proceedings of the National Academy of Sciences, vol. 116, 2019, pp. 4156–4165.
[33] D.-H. Lee, et al., Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks, in: Workshop on Challenges in Representation Learning, ICML, vol. 3, 2013, pp. 896.
[34] J. Lienen, E. Hüllermeier, From label smoothing to label relaxation, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, 2021, pp. 8583–8591.
[35] J.C. Luers, M. Grosheva, M. Stenner, D. Beutner, Sialoendoscopy: prognostic factors for endoscopic removal of salivary stones, Arch. Otolaryngol–Head Neck Surg. 137 (2011) 325–329.
[36] H. Ma, D. Zeng, Y. Liu, Learning optimal group-structured individualized treatment rules with many treatments, J. Mach. Learn. Res. 24 (2023) 1–48.
[37] D.J. MacKay, Information Theory, Inference and Learning Algorithms, Cambridge University Press, 2003.
[38] O.J. Maclaren, R. Nicholson, What can be estimated? identifiability, estimability, causal inference and ill-posed inverse problems, arXiv preprint arXiv:1904.02826, 2019.
[39] F. Marchal, A.-M. Kurt, P. Dulguerov, W. Lehmann, Retrograde theory in sialolithiasis formation, Arch. Otolaryngol.–Head Neck Surg. 127 (2001) 66–68.
[40] M. McGurk, M. Escudier, J. Brown, Modern management of salivary calculi, J. Br. Surg. 92 (2005) 107–112.
[41] H. Meng, Y.-Q. Zhao, H. Fu, X. Qiao, Near-optimal individualized treatment recommendations, J. Mach. Learn. Res. 21 (2020) 1–28.
[42] E. Miranda, A survey of the theory of coherent lower previsions, Int. J. Approx. Reason. 48 (2008) 628–658.
[43] E. Miranda, S. Destercke, Extreme points of the credal sets generated by comparative probabilities, J. Math. Psychol. 64 (2015) 44–57.
[44] A. Montanari, B.N. Saeed, Universality of empirical risk minimization, in: Conference on Learning Theory, PMLR, 2022, pp. 4310–4312.
[45] N. Natarajan, I.S. Dhillon, P.K. Ravikumar, A. Tewari, Learning with noisy labels, Adv. Neural Inf. Process. Syst. (2013) 26.
[46] H. Parikh, C. Varjao, L. Xu, E.T. Tchetgen, Validating causal inference methods, in: International Conference on Machine Learning, PMLR, 2022, pp. 17346–17358.
[47] M. Sajjadian, R.W. Lam, R. Milev, S. Rotzinger, B.N. Frey, C.N. Soares, S.V. Parikh, J.A. Foster, G. Turecki, D.J. Müller, et al., Machine learning in the prediction of depression treatment outcomes: a systematic review and meta-analysis, Psychol. Med. 51 (2021) 2742–2751.
[48] M. Schapher, M. Koch, D. Weidner, M. Scholz, S. Wirtz, A. Mahajan, I. Herrmann, J. Singh, J. Knopf, M. Leppkes, et al., Neutrophil extracellular traps promote the development and growth of human salivary stones, Cells 9 (2020) 2139.
[49] S. Schrøder, P. Homøe, N. Wagner, A. Bardow, Does saliva composition affect the formation of sialolithiasis? J. Laryngol. Otol. 131 (2017) 162–167.
[50] R. Shwartz-Ziv, A. Armon, Tabular data: deep learning is not all you need, Inf. Fusion 81 (2022) 84–90.
[51] C. Suhler, P. Churchland, Psychology and medical decision-making, Am. J. Bioeth. 9 (2009) 79–81.
[52] P. Tigas, Y. Annadani, A. Jesson, B. Schölkopf, Y. Gal, S. Bauer, Interventions, where and how? experimental design for causal models at scale, Adv. Neural Inf. Process. Syst. 35 (2022) 24130–24143.
[53] J.E. Van Engelen, H.H. Hoos, A survey on semi-supervised learning, Machine Learning 109 (2020) 373–440.
[54] V. Vapnik, The Nature of Statistical Learning Theory, Springer Science & Business Media, 2013.
[55] Y. Yang, Z. Xu, Rethinking the value of labels for improving class-imbalanced learning, Adv. Neural Inf. Process. Syst. 33 (2020) 19290–19301.
[56] J. Yoon, J. Jordon, M. Van Der Schaar, GANITE: estimation of individualized treatment effects using generative adversarial nets, in: International Conference on Learning Representations, ICLR 2018, 2018.
[57] Z.-H. Zhou, A brief introduction to weakly supervised learning, Natl. Sci. Rev. 5 (2018) 44–53.