A Giant with Feet of Clay: On the Validity of the Data that Feed Machine Learning in Medicine

Federico Cabitza, Davide Ciucci and Raffaele Rasoini

Abstract This paper considers the use of machine learning in medicine by focusing on the main problem that it has been aimed at solving or at least minimizing: uncertainty. However, we point out how uncertainty is so ingrained in medicine that it biases also the representation of clinical phenomena, that is the very input of this class of computational models, thus undermining the clinical significance of their output. Recognizing this can motivate researchers to pursue different ways to assess the value of these decision aids, as well as alternative techniques that do not "sweep uncertainty under the rug" within an objectivist fiction (which doctors can come up by trusting).

Keywords Decision support systems · Machine learning · Uncertainty

1 Motivations and Background

It is a truism to say that uncertainty permeates contemporary medicine—not much differently than it has always been—as it has also been confirmed by extensive studies in the field of sociology and medicine itself (e.g., [20, 49]). We cope with some form of uncertainty when we cannot pinpoint a phenomenon exactly or when we cannot

An extended version of this paper can be found on the arXiv platform at the following address: https://arxiv.org/abs/1706.06838.

F. Cabitza · D. Ciucci (✉)
Università degli Studi di Milano-Bicocca, Milan, Italy
e-mail: ciucci@disco.unimib.it

F. Cabitza
IRCCS Istituto Ortopedico Galeazzi, Milan, Italy
e-mail: federico.cabitza@unimib.it

R. Rasoini
IRCCS Don Gnocchi Foundation, Milan, Italy
e-mail: raffaele.rasoini@tiscali.it

F. Cabitza et al. (eds.), *Organizing for the Digital World*, Lecture Notes in Information Systems and Organisation 28, https://doi.org/10.1007/978-3-319-90503-7_10

measure it precisely (i.e., approximation, inaccuracy); when we do not possess a complete account of a case (incompleteness, inadequacy); when we cannot predict what it will come next (unpredictability for randomness or excessive complexity); when our observations seem to contradict each other (inconsistency, ambiguity); and, more generally, when we are not confident of what we know. In clinical practice, all of these phenomena occur on a daily basis, several times. Medical doctors can be uncertain on almost every aspect of their practice: on how to classify patients' conditions (diagnostic uncertainty); why and how patients develop diseases (etiological u., pathophysiological u.); what treatment will be more appropriate for them (therapeutic u.); whether they will recover with or without a specific treatment (prognostic u.), and so on. In this picture, technology has often been proposed—and seen—as a solution. In the words by Reiser [44, p. 18]: *From the beginning of their introduction in the mid-nineteenth century, automated machines that generated results in objective formats [...] were thought capable of purging from health care the distortions of subjective human opinion [and] to produce facts free of personal bias, and thus to reduce the uncertainty associated with human choice.*

Clearly, also computing technology has been proposed to address all of the above areas of uncertainty—to either control or minimize it: the first computational support, what was then called a rule-based expert system, was introduced more than 40 years ago to propose a "quantification scheme which attempts to model the inexact reasoning processes of medical experts" [48].

After the introduction of many and different computational systems, a new class of applications has recently emerged in the health care debate: the *decision support systems* embedding predictive models that have been developed by means of *machine learning* methods and techniques. These systems, which for the sake of brevity we will call ML-DSS, have recently raised a strong interest among the medical practitioners of almost every corner of the world in virtue of their high accuracy at an unprecedented extent [19, 23]. This is reflected by the stance of commentators that have recently shared their thoughts from some of the most impacted journals of the medical community (e.g. [32, 38]). These voices do not clearly indulge in techno-optimistic claims and do not refrain from offering some words of caution [11]; however, the recent successes of ML-DSS in medical imaging pose the issue of how these systems and their improved versions will impact in the near future those medical professions whose tasks are mostly based on pattern recognition, like radiologists, pathologists and dermatologists and how it will impact health care in general [32, 38]. In regard to this, two elements should be object of further scholarly interest and research, which are bound together by a feedback loop making their mutual influence subtle but yet hard to pinpoint. First: how ML-DSS can bias human interpretation and decision, or *automation bias*. Second: how human interpretation and classification can bias the ML-DSS performance, or *information bias*. While the former case is still neglected but some first studies are shedding light on it [39], in this paper we will focus on the latter case, which is almost completely ignored, especially by the computer scientists and designers of ML models. Nevertheless

information bias, which we will define in the next section, regards the quality[1] of both the training and input data of ML-DSS, and hence has got the potential to undermine the reliability of the output of ML-DSS. Our point is that a renovated awareness of the irreducible nature of the uncertainty of medical phenomena, even in regard to their plain representation into medical data, can help put in the right perspective the current potential of ML-DSS and motivate the exploration of alternative ways to conceive them and validate their indications.

2 Information Bias, the Open Secret of Medical Records

Before considering information bias from a medical perspective, let us recall what a ML predictive model is. A predictive model, no exception those developed with a machine learning approach,[2] are *functionally relational* models that bind input data to one category out of a set of predefined ones (most of the times encompassing only two options, like positive/negative). This category is the output (or prediction) of the model. To this aim, the model is progressively fine-tuned on the basis of what ML experts call *experience* [37, p. 2], that is input data that have been already classified in terms of a specific category. In the case of medical classification (for both diagnostic or prognostic aims) the above "experience" is a set of cases that have been already classified "correctly" *according to some gold standard method*. In so doing, the machine can learn the model, that is the hidden relationship between the cases (as long as they are represented in terms of the same attributes and characteristics), and hence their correct interpretation. Grounding on this model, the ML-DSS can "predict" the right category or label when fed with new cases, as long as these are sufficiently similar to those ones with which it has been trained.

The point we address here is: how much valid and reliable is the above "experience" on which ML-DSS learn their predictive model? Here it comes the concept of information bias [2], and the related one of information variability, which both undermine the extent we can be certain of the available data, and hence of the predictions ML models can infer from them. Information bias is a collective name for all the human biases that distort the data on which a decision maker (or a computational decision support) rely on, and that account for the validity of data, that is the extent these represent what they are supposed to represent accurately.

This kind of bias[3] can take various forms including, most manuals concur, measurement error, misclassification and miscoding. However, this bias should *not* be only associated with errors and mistakes by the physicians due to either negligence,

[1] This is a vague term: here we mean data quality mainly in terms of accuracy and validity.

[2] In what follows we introduce the concept of ML predictive model with reference to supervised discriminative (or classification) models, by far the most frequently used in medicine.

[3] While biases are, strictly speaking, mental prejudices, idiosyncratic perceptions and cognitive behaviors producing an either impairing and distorting effect, here we rather intend the effect (by metonymy), that is the "error" in the data recorded and the decisions taken caused by the bias itself.

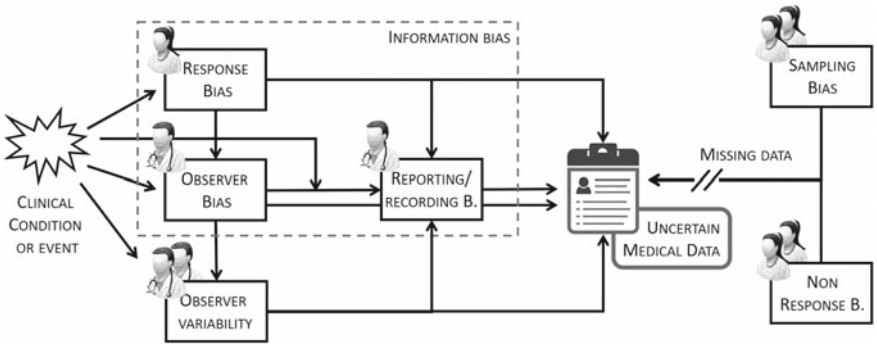

Fig. 1 The main biases affecting the validity and reliability of medical data. Sampling and non-response biases are indicated to account for the lack of information that, if present, could reduce uncertainty of representation

incompetence or inexperience. In medicine, information bias can be due to both patients and physicians in different but intertwined ways (see Fig. 1). Patient can (often unaware) contribute in terms of *response bias*, that occurs whenever what the patient says or reports is inaccurate, incomplete or both. This bias occurs when patients either exaggerate or understate their conditions (for many reasons and often in good faith), or whenever they intentionally suppress some information (e.g., like in case of sexually transmitted diseases or drug history for the related social stigma) or distort it or when their recall is limited or flawed, or simply because they do not understand what physicians ask them or they aim to respond how (they believe) physicians expect them to (cf. the particular kind of response bias known as "social desirability bias"). A large extent of response bias can be related to the inability of the observer to get confidence of the respondent. As an example, a recent study [47] focused on the degree of concordance agreement between patients and cardiologists in regard to the presence of angina pectoris and its frequency. The study showed that when patients reported monthly angina symptoms, cardiologists agreed only 17% of time, while among patients with angina symptoms reported daily/weekly, approximately one quarter of them were noted as having no angina by their physicians. Besides the above mentioned condescending bias, it is well known that patients can exhibit behaviors (cf. Hawthorne effect), or even levels of some physiological parameter (usually blood pressure in what is also known as 'white-coat effect') when they are under examination in a clinical setting that they would not exhibit in other settings.

Physicians introduce *observer bias* in their data due to both perceptual, cognitive and behavioral traits, weaknesses or just "bad habits"; this bias also occurs whenever they favor one type of response or measurement over others (cf. confirmation bias). Those who observe a clinical condition are often those who report it in the medical record. In this case observer bias can blur with is denoted as either recording, reporting, or coding bias. In particular, this latter distorting factor can be traced back to many causes, from the least common and most poorly studied, like digit preference and conflictual coding, to the most pervasive ones, like the intrinsic inadequacies of

any classification schema. Conflictual coding can affect the accuracy and completeness of medical data when proper reporting clashes with the personal interests of those who are supposed to document a clinical condition (like in the case of blood pressure recordings within a quality and outcome assessment framework of incentives [13]). Digit preference occurs when measurements are more frequently recorded ending with 0 or 5, or as results of arbitrary rounding off. In [27] the authors observed a much larger occurrence of these two digits in renal cell carcinoma measurement and concluded that this recording behavior could affect the determination of tumour stage, "with resulting consequences in regard to prognosis and patient management". Moreover, coding variability that leads to a lack of reliability can happen even when instructions on how to proper code are well known: a study [3] compared consistency of coding supposedly clean and high-quality data-source like clinical research form from observational studies among three professional coders, each using the same terminology and with the same instructions. All three coders agreed on the same core concept 33% of the time; two of the three coders selected the same core concept 44% of the time; and, no agreement among all three was found 23% of the time. Moreover, no significant level of agreement beyond that due to chance was found among the experts. On the other hand, the shortcomings of classifying taxonomies and measurement scales would deserve a study of its own. In a famous work, Star and Bowker [6, p. 69] hinted at some of these inadequacies, which include: temporal rigidity; a one-size-fits-all nature in regard to meaning and implications; and, as also discussed in [52], the reflection of disciplinary interests, agendas and priorities.

Last, but not least, information bias in medicine can also be traced back to some sort of *intrinsic ambiguity* of the medical conditions being documented, due to either their instability over time, or to variability across subjects and across observers. A noteworthy example of this sort of ambiguity can be found in a recent study by Dharmarajan and colleagues [16]. This study focused on elderly people diagnosed (at hospital admission) with one of the following conditions: pneumonia, chronic obstructive pulmonary disease, or heart failure. These are three common conditions of the elderly that are responsible for breathlessness and other warning symptoms usually requiring hospital admission. The authors showed that patients regularly received, during hospital stay, concurrent treatment for two or more of the above cardiopulmonary conditions and not only for the main diagnosis identified at hospital admission. This exemplifies the fact that in real-world clinical practice, patients' clinical pictures are often blurry and not capable of being associated with clear-cut labels as expressed instead in textbooks and clinical practice guidelines. Indeed, even common clinical syndromes have disease presentations that often fall in-between traditional diagnostic categories. The common and relevant overlap of medical treatments as in the case mentioned above highlights the intrinsic ambiguity of clinical phenomena and the downstream uncertainty that medical doctors face in choosing what they deem a single right therapeutic strategy for a specific disorder.

This latter sort of variability, as well as the biases mentioned above cannot be addressed by improving the accuracy of any measurement tool, or by any other contrivance conceived from the engineering standpoint. Moreover, the extent these biases are expressed in a clinical setting varies a lot: although they look as abstract

and general categories, biases are always exhibited by someone in particular, they are highly situated, and depend on personal skills, like clinical perspicuity, life-long acquired competencies, and contingent workloads. Since the impact of personal biases are difficult (if impossible) to prevent, medical organizations try to minimize them with redundancy of effort, like relying on double checking and on second (or multiple) opinion. However, multiple opinions are both a resource to fight biases (by averaging multiple observations and measures), and, paradoxically enough, a source of low reliability and further uncertainty ('Quot capita, tot sententiae'). Indeed, when more than one physician are supposed to determine the presence of a sign, make a diagnosis, or assess the severity of a condition, *observer variability* is introduced to account for the discrepancies in their decisions and for any difference in considering the same conditions. Observer variability has to do with the reliability of the judgment of so called medical "raters", and with the agreement that these latter ones reach independently of each other when they either measure, classify or interpret the *same* phenomenon (e.g. an electrocardiogram, a radiography, a pathological sample, etc.) to make a decision, mainly a diagnosis.

In medicine, not only multiple raters could classify the same phenomenon in different ways, but also the same doctor can disagree with herself, examining the same case after a certain amount of time, or in different environment conditions (e.g., with respect to workload, interruption rate, work shift).

To account for the extent the majority decision (in case of multiple raters) or the category chosen more frequently can be considered reliable, and hence "true", the so called *inter-rater agreement* (or reliability, IRR) is measured.

3 Between Gold Standards and Ghost Standards

A "gold standard" (or with a less evocative but more correct expression "criterion standard") is a reference method to ascertain medical truth in regard to the accuracy of any diagnostic test. By 'reference' here we mean the 'best one' under reasonable conditions, that is the method that 'by definition' is capable to pose the so called "ground truth" for any practical aim, including the development of a ML-DSS. However, the degree of truth that a gold standard usually reaches is far from resembling an accurate, unambiguous and unique representation of medical facts that computer scientists long for their "ground-truthing", i.e., the process of gathering objective data to train a ML model. In fact, even in regard to those tests that are usually considered the most reliable and definitive gold standards, like post-mortem, histological and genetic examinations, whenever there is a human factor, variability, and hence uncertainty, can emerge [7, 55, 56] as if the observers were called to observe and account for phantom phenomena. In all of these cases, IRR scores can give an idea of the extent the data that doctors collect, which glitter in medical datasets, are golden or alloyed.

Medical researchers use several techniques to measure IRR: the most frequently used is the Cohen's Kappa, although this is applicable only to categorical values

assigned by two raters. Recently also the Krippendorf' Alpha has found some application, probably for its known advantages on the kappa, like the capability to address any number of raters (not just two), values of any level of measurement (i.e., categorical, ordinal, interval, etc.), and datasets with missing values, which are very common in medical records.

All these metrics are intended to assess the degree of agreement *beyond chance* (that is considering the fact that raters can agree not only because they believe the same thing, but also by chance). As such, there is a lot of controversy on their validity (lacking any model of how chance affects the rater decisions, let alone of the different ways to *misinterpret* a phenomenon) and, above all, regarding how to interpret their scores [34].

To give an example of the above concepts, we consider an ambit where ML-DSS have recently reached levels of diagnostic accuracy (at least) on a par with human doctors. A convolutional neural network was recently trained and then evaluated in regard to the detection of diabetic retinopathy (which is cause of 1 case of blindness out of 10) in a wide dataset of retinal fundus photographs [23]. In this study, the authors reported high levels of both sensitivity and specificity for their ML-DSS according to the gold standard that they decided to adopt, i.e. the majority decision of a panel of board-certified ophthalmologists analyzing the same retinal fundus photographs. As a matter of fact, some authors reported that the adoption of this gold standard can be considered controversial [59], since this may compare unfavorably with other gold standards used in previous studies on diabetic retinopathy (i.e. standardized centralized assessment of images or optical coherence tomography). In fact, the prevalence of diabetic retinopathy may vary significantly whether this condition is evaluated through monocular fundus photographs or, rather, through optical coherence tomography [58]. This could turn out to be relevant since diagnostic accuracy metrics are dependent on the prevalence of diseases according to the Bayes' theorem. Thus, two questions are at stake here. On the one hand, whether similar successful results would have been obtained if a different gold standard (e.g., optical coherence tomography) had been used. On the other hand, even if we assume eye fundus photographs as an indisputable, unique and reliable gold standard for the diagnosis of diabetic retinopathy, IRR among ophthalmic care providers has been shown to be very low (i.e., kappa between 0.27 and 0.34 for different diagnostic analyses); and still inadequate, even if higher, among retina specialists (kappa between 0.58 and 0.63) [46].

4 Garbage In, Gospel Out

The question of the quality of medical record and of the data extracted from there is still understudied [9]. The assumption that medical data could support secondary uses has been challenged since almost 25 years ago, and also strongly so, e.g., by Reiser [45], who described several cases of erroneous, missing and ambiguous data, and by Burnum [8], who provocatively wrote that "all medical record information

should be regarded as suspect; much of it is fiction" (p. 484)", and that the introduction of health information technology had not led to improvements in the quality of medical data recorded therein, but rather to the recording of a greater quantity of "bad data".[4] In those same years, van der Lei was among the first ones to warn against the reuse of clinical data for other goals other than care and proposed what since then is known as the first law of informatics: '[d]ata shall be used only for the purpose for which they were collected' [54]. These claims are reflected in some recent research articles in higly impacted medical journals that warn about the risks and challenges related to the use of routinely collected data (e.g. from electronic health records) for clinical decision making [1, 28].

In light of the phenomena of both low quality and uncertainty that are intrinsic to the production of medical data, what are the main implications for the machines that are fed with this information? As widely known, many factors can contribute in downsizing the performance of a ML-DSS. Just to mention a few that we observed in the hospital domain: the fact that medical data seldom meet the common assumptions that training data should ideally possess: their attributes are seldom independent and identically distributed (IDD); their distribution is not uniform or normal; missing data do not occur randomly (in fact they often indicate an either good or just better health condition that relieves practitioners from the need to record it with continuous effort); data can be strongly unbalanced with respect to the number of healthy and positive cases, or to the real incidence of a pathological condition; they are not temporally stable (for instance, computer interpretations of electrocardiograms recorded just one minute apart were found significantly different in 4 of 10 cases in [50]); they can fall short of representing the target population (*sampling bias*) or to make explicit any potential confounding variable (especially those related to "external" medical interventions [41]).

In this view, misclassified cases by information bias could be seen as just another issue to cope with. However, our point is that to consider misclassification a defect of data collection is a conceptual error as long as it is considered a *mis*-classification: as we saw above, it is just a classification where independent observers disagree and classify the same phenomenon differently, to the best of their competence, perspicacity and perceptual acuity.

This variability if often neglected even by doctors, and few studies indulge in reporting low IRR scores. No wonder then that the related uncertainty is dispelled as closely as possible to the "source", as also the official guidelines for medical coding and reporting (International Classification of Diseases, Ninth Revision, Clinical Modification, ICD9-CM) ratify in an explicit way: *"If the diagnosis documented at the time of discharge is qualified as 'probable', 'suspected', 'likely', 'questionable', 'possible', or 'still to be ruled out', or other similar terms indicating uncertainty, code the condition as if it existed or was established"* [43, p. 90]. Alternatively uncertainty is sublimated in the (statistically significant) consensus of a sufficiently wide group of experts [51].

[4]Moreover, Burnum traced back this lie of the land to "standards of care and a reimbursement system [that is] blind to biologic diversity".

Adopting different gold standards could affect ML-DSS significantly. We illustrate this by mentioning the case of Carpal Tunnel Syndrome (CTS): this is a kind of functional hand impairment that is frequently observed due to the compression of the median nerve at the wrist. This syndrome is commonly diagnosed and often referred to surgical treatment through either the sole physical examination by *orthopedic surgeons* or by a nerve conduction examination (electromyography) by *neurologists* [4]. In the last years, alternative diagnostic methods have been proposed to improve diagnostic accuracy for CTS, like the ultrasonography of the median nerve of the arm. These tests which have shown different results in accuracy metrics when compared to the previous standards mentioned above (i.e. physical examination or electromyography) [4]. These diagnostic divergences, if neglected in the training of a ML classifier aimed at the diagnosis of CTS [36], may result in the ossification into the model of an arbitrarily partial version of the ground truth (that is whether patient X is really affected by CTS or this syndrome can be ruled out) and hence to unpredictable downstream *clinical* consequences. For instance, it has been observed that a number of patients diagnosed with CTS who had undergone surgery did not receive any relevant benefit from the invasive treatment, and that this could be explained in terms of wrong upstream diagnoses [21]. A ML-DSS that has learned the *uncertain* (i.e., right for a standard, wrong for another) mapping between the patient's features and one single diagnosis will propose its advice within a dangerous "close-world assumption" (that is: all the relevant features have been considered and the mapping between the input and the desired output is acquired as accurate and reliable), which is never challenged *by design*.

On the other hand, in the open world of hospital wards physicians are used to observer variability and less-than-perfect gold standards whenever they consult the medical data that are produced by their colleagues and even by themselves. Conversely, designers of computational systems usually do not consider the case that the input of their system can be inherently and irremediably biased and inaccurate (to some extent), they assume it true. The primary concern of ML-DSS designers are the completeness, timeliness and consistency [9] of the datasets that they feed into the machines. There is little (if any) recognition that medical data could not be any better than "dirty" data with which to think to optimize a ML model adequately would be highly optimistic if not over-ambitious. Contrarian thinking would then suggest to look with some caution at the high accuracy rates that are often reported in the specialist literature (e.g. in [19]) even assuming that model overfitting has been duly avoided.

This is hardly considered when medical data are taken from the context where they have been natively produced to support coordination, knowledge sharing, sense making and decision making and they are transformed into data sets to feed in some computational systems. Neglecting the gap between the primary use of medical data (i.e., care) and any secondary use (e.g., ML-DSS training) could mislead those who have to design trustworthy decision support systems, and also probably jeopardize the actual improvement of the ML-DSS performance on new and real data other than the training data.

This points to the difference between *research data*, which are usually used for ML-DSS training and optimization; and *real-world* data, which are produced in real-life clinical situations. While research data are not made up on purpose to get high accuracy, they are nevertheless selected, cleansed, and *engineered* to an extent that is completely unrealistic or unfeasible to replicate in actual clinical settings. This is not only a matter of generalizability and interpretability of the model. It is also a matter of different ways to evaluate ML-DSS. The most common one can be considered *essentialist* [12], in that it focuses on accuracy and other performance measures (like F1-scores and AUC-ROC) that are appraised in a laboratory setting (i.e., *in laboratorio*). An alternative and still neglected approach, which is the *consequentialist* one, focuses on the actual clincal outcomes (consequences) produced in situated practice (i.e., *in labore*), that is in the original context of work of the physicians involved and in their actual relationship with patients, when decisions must be converted into real-life choices that must align with the patients' attitudes, preferences, fears and hopes, as well as with the economic feasibility of the available options.

5 Embracing Uncertainty, Also in Computation

As hinted above, there are many types of uncertainty in medicine, which affect medical records in different ways. For a certain attribute (i.e., variable) that is pertinent for a certain case, users could ignore what value is applicable, let alone true; or what single value is true among a finite set of values that are known to be equally applicable. Users could be uncertain between two values from the above set, or among many. Moreover, they could prefer some options with respect to others. If single users are certain about a value, they could nevertheless disagree among each other (and even with themselves over time). Ultimately, they could be uncertain among different values at various levels of confidence with respect to each other (e.g., in a dichotomous domain, which is the simplest, doctor A is *fairly* certain that the condition is pathological, doctor B is *strongly* certain).

As shown by Svensson et al. in [51], the performance of ML-DSS is negatively impacted and deeply undermined when fed with medical datasets that are intrinsically uncertain. Their idea is to employ conventional statistical tests to reduce the variability of the data produced by different observers by choosing the values that have been proposed by a statistically significant majority of the observers. However effective, this could be also seen as a way to discard the richness of a multi-value representation that accounts for a manifold phenomenon, which competent and skillful observers can describe each in her own, and partially sound yet specifically irreproachable, way.

Thus, if we take the "dirtiness" and "manifoldness" (seen as sides of the same coin) of medical data as a given factor of medicine, one could wonder: how can ML techniques take these constraints seriously, and even exploit them to get a richer picture of the phenomena of interest; and hence build models that could really support human experts in their daily, and uncertain, practice?

First of all, let us remark once more that there exist different kinds of uncertainty, as exemplified above. Indeed, if we look at the different classifications and taxonomies of uncertainty [40], we find a long list of terms, such as: Absence, Ambiguity, Approximation, Belief, Conflict, Confusion, Fuzziness, Imprecision, Inaccuracy, Incompleteness, Inconsistency, Incorrectness, Irrelevance, Likelihood, Nonspecificity, Probability, Randomness, . . . Each form has its own tools to represent and manage it, to name a few: probability theory, fuzzy sets, possibility theory, evidence theory, rough sets. By large, the predominant role is played by probability theory, and machine learning is not an exception in this attitude. However, there exist solutions (in some case preliminary attempts) to incorporate other tools in machine learning (see e.g., [5, 15, 30]). There are several reasons why these approaches are not well established in ML, as widely discussed in [31] for the fuzzy set case: sometimes the new tools are naïves; there is not a connection among different communities; there is a problem of credibility for many young, or at least not as well established as probability theory, disciplines. However, if we want to deal with all the different forms of uncertainty, it is needed and possible to directly manage them. Indeed, it is our belief that the above discussed flaws in data can be addressed in machine learning by making use of the appropriate tools. In the following we give some hints on how this can be done, making reference to the biases previously discussed.

At first, let us consider the problem of representing a rater reliability. It is always assumed that ratings are exact, though they may be classified as "deterministic" or "random", where random means that "the rater is uncertain about the response category" [24]. More than a question of randomness, this description points to a form of epistemic uncertainty which can be handled by not assuming exactness and introducing graduality on the judgment scale of a rater. For instance, we could have three levels of certainty (i.e., low/fair/good) on the assigned score and/or the rater can express her uncertainty by selecting more than one score with its own level of certainty. This kind of uncertainty can be applied also to the input data and we can represent the fact that a patient has *low* headache and *high* nausea, whereas in a dichotomic situation we were forced to say *no* headache and *yes* nausea. This situation can be handled with Possibility Theory and, in particular, with its simplified form of certainty-based model [42], which is more interpretable and simple from a computational standpoint. Of course, machine learning tools have to be modified to comply with this model, though some steps in this direction already exists [25, 29]. Worth mentioning is also the fuzzication of the Arden syntax aimed "at simplifying programmes which process indeterminate data by means of fuzzy logic" [57].

As another problem, let us consider a numerical information, such as the one obtained by some measurement. Of course, any point value brings an imprecision, due to the instrument itself or to an average among repeated measures, etc. Thus, we can consider to represent the information in form of an interval and directly work with it. To this scope, interval arithmetics and fuzzy arithmetics [35] give the formal instruments to operate with this kind of representation.

Finally, it is well known that data often come with missing values. A standard approach in ML is to impute them in order to get a complete dataset. Of course, in this way, we loose the original information and some errors or at least imprecisions

are introduced. On the contrary, we should not get rid of the missing values and moreover take into account that a missing value can have several meanings, such as ignorance, non existence, etc. Rough set theory includes rule induction methods that comply with missing values and also with different meanings of it. In particular, we point the attention to the works by J. Gryzmala-Busse and his MLEM algorithm [22]. We also notice that some attempts to directly deal with missing values exist also in other ML approaches, such as fuzzy clustering [26].

6 Conclusions

Fox [20] in her relevant work on the sociology of medical knowledge once wrote that uncertainty has become the hallmark of the entire field of medicine. For this reason, confronting uncertainty has been the first and foremost driver for the introduction of computational Decision Support Systems in medicine and their increasingly wide adoption in clinical settings. We could just speculate on why medicine has turned to technology to "make sense of health data" (to cite a position paper on Nature published a couple of years ago [18]). Quite subtly, Katz [33] has argued that the traditional mechanisms that physicians use to adopt to cope with uncertainty (e.g., terminological standards, standard care protocols, guidelines based on statistical studies) can slowly push them towards disregarding or even opposing uncertainty.

Irrespective of the root causes of this situation, the digitization of medicine has contributed to shifting the idea of uncertainty, from being a natural and irreducible element of medical practice [49] in the interpretation of subtle and sometimes contradictory clues in the existential and complex context of idiosyncratic patients, into the domain of those rational problems that can be modeled to pursue an engineering solution, or even a computational one.

In this paper we have briefly explored the blurring boundaries between what computer scientists and medical doctors pursue in medical data: data accuracy and completeness the former ones; trustworthiness and meaningfulness the latter ones [9]. We have also shed light on information bias and observer variability, which separate us from getting an absolute true, universal and reliable *representation* of a physical (let alone psychological or mental) phenomenon. In particular for the ML designers, we have pointed out that information bias does not regard only the labelling of data set, i.e., the information on which a ML-DSS is trained to predict other labels accurately; but it also (and above all) affects the whole input data, in both training and prediction, especially in regard to nominal and ordinal variables.

In light of these different viewpoints, we outline a couple of recommendations along the general framework by Domingos, who conceives ML problems as a combination of *representation*, *optimization* and *evaluation* [17]. From the representation perspective, computer scientists should not settle for "polished data" but rather "get to the source" of medical data: the multiple, possibly divergent opinions of experts. This means to be wary of researches where the gold standard is not reported or it is a dataset annotated by a single, or just a couple of physicians. Moreover, if the

adopted gold standard is based on a consensus reconciliation of divergent opinions, the authors of those researches should also be aware that they proceed considering all of these divergences plain mistakes. If they are less than certain this is fair, they should offer a word of caution on the potential arbitrariness of the clearly-cut classification they have used in their study. In the study design phase authors could also ask the competent observers the degree of self-perceived confidence with which they share their ideas and produce their data. This ordinal scale could be used to weight the multiple values of a single representation, so that the ML algorithms can leverage again the knowledge of the domain experts to build a coherently *fuzzy* representation. Furthermore they could annotate the representativeness of each value in terms of a three-way partitioning (e.g, belonging to the majority opinion, belonging to the minority, belonging to neither with statistical significance [10]). In any case, ML researchers should always report how they did collect their ground truth, relying on what gold standard.

In regard to optimization, further research should be devoted in transferring techniques and methods from the rough set theory [53] domain into the ML arena.

In regard to evaluation, the ball could be passed to the medical practitioners again. They should develop a wariness of any essentialist evaluations of ML-DSS performance that are carried out *in laboratorio*, on *research data* and are expressed as *accuracy metrics*. Rather, they should demand to the ML-DSS designers (and their advocates) *evidence-based* validations of their systems, that are focused on *health outcomes* and adopt them only once some further information has been given about, e.g. the trade-off between the internal (i.e., bias) and external (i.e., variance) validity of the model (regarding also the extent the ML-DSS could fit multimorbid cases, instead of being excessively specialized for one disease); and between its prediction power and its interpretability [14], that is its *scrutability* by doctors and lay users to understand why the ML-DSS has suggested them a certain decision over possible others and make the "hybrid" agency of man-and-machine more accountable towards the colleagues, the patients and their dears. Even more than that, ML-DSS should be object of a *value-based* assessment, where researchers invest time and effort on the evaluation of their systems in the mid- long-term after their deployment in real settings and their appraisal is conducted in terms of user and patient satisfaction, in terms of effect size on clinical outcomes, and eventually in terms of cost reduction or better service provision. All these elements should not be overlooked or given for granted, especially in light of the perils of automation bias (such as deskilling, technology overreliance and overdependence) not least the surreptitious increase of trust by doctors in numbers and the "objective facts" (cf. McNamara fallacy) that the reckless application of *machine learning* in response to an excessive *human yearning* for certainty could bring in, especially in fields where this is likely to turn out to be only a dream of ignorance.

References

1. Ahmad, F.S., Chan, C., Rosenman, M.B., Post, W.S., Fort, D.G., Greenland, P., Liu, K.J., Kho, A., Allen, N.B.: Validity of cardiovascular data from electronic sources: the multi-ethnic study of atherosclerosis and HealthLNK. Circulation 117 (2017)
2. Althubaiti, A.: Information bias in health research: definition, pitfalls, and adjustment methods. J. Multidiscip. Healthc. **9**, 211 (2016)
3. Andrews, J.E., Richesson, R.L., Krischer, J.: Variation of SNOMED CT coding of clinical research concepts among coding experts. J. Am. Med. Inf. Assoc. **14**(4), 497–506 (2007)
4. Bachmann, L.M., Jüni, P., Reichenbach, S., Ziswiler, H.R., Kessels, A.G., Vögelin, E.: Consequences of different diagnostic gold standards in test accuracy research: Carpal tunnel syndrome as an example. Int. J. Epidemiol. **34**(4), 953–955 (2005)
5. Bello, R., Falcon, R.: Rough Sets in Machine Learning: a review, pp. 87–118. Springer International Publishing, Cham (2017)
6. Bowker, G.C., Star, S.L.: Sorting Things Out: classification and its consequences. MIT press (2000)
7. Braun, R., Gutkowicz-Krusin, D., Rabinovitz, H., Cognetta, A., Hofmann-Wellenhof, R., Ahlgrimm-Siess, V., Polsky, D., Oliviero, M., Kolm, I., Googe, P., et al.: Agreement of dermatopathologists in the evaluation of clinically difficult melanocytic lesions: how golden is the gold standard ? Dermatology **224**(1), 51–58 (2012)
8. Burnum, J.F.: The misinformation era: the fall of the medical record. Ann. Int. Med. **110**(6), 482–484 (1989)
9. Cabitza, F., Batini, C.: Information quality in healthcare. In: Data and Information Quality, Chap. 13, pp. 421–438. Springer (2016)
10. Cabitza, F., Ciucci, D., Locoro, A.: Exploiting collective knowledge with three-way decision theory: cases from the questionnaire-based research. Int. J. Approx. Reason. **83**, 356–370 (2017)
11. Cabitza, F., Rasoini, R., Gensini, G.F.: Unintended consequences of machine learning in medicine. Jama **318**(6), 517–518 (2017)
12. Cappelletti, P.: Appropriateness of diagnostics tests. Int. J. Lab. Hematol. **38**(S1), 91–99 (2016)
13. Carey, I., Nightingale, C., DeWilde, S., Harris, T., Whincup, P., Cook, D.: Blood pressure recording bias during a period when the quality and outcomes framework was introduced. J. Hum. Hypertens. **23**(11), 764 (2009)
14. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., Elhadad, N.: Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1721–1730. ACM (2015)
15. Denœux, T., Kanjanatarakul, O.: Evidential Clustering: a review, pp. 24–35 (2016)
16. Dharmarajan, K., Strait, K.M., Tinetti, M.E., Lagu, T., Lindenauer, P.K., Lynn, J., Krukas, M.R., Ernst, F.R., Li, S.X., Krumholz, H.M.: Treatment for multiple acute cardiopulmonary conditions in older adults hospitalized with pneumonia, chronic obstructive pulmonary disease, or heart failure. J. Am. Geriatr. Soc. **64**(8), 1574–1582 (2016)
17. Domingos, P.: A few useful things to know about machine learning. Commun. ACM **55**(10), 78–87 (2012)
18. Elliott, J.H., Grimshaw, J., Altman, R., Bero, L., Goodman, S.N., Henry, D., Macleod, M., Tovey, D., Tugwell, P., White, H., et al.: Informatics: make sense of health data. Nature **527**, 31–32 (2015)
19. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(7639), 115–118 (2017)
20. Fox, R.C.: Medical uncertainty revisited. Handb. Soc. Stud. Health Med. 409–425 (2000)
21. Graham, B.: The diagnosis and treatment of carpal tunnel syndrome: surgerywhether open or closed works, but only if the diagnosis is right. BMJ. Br. Med. J. **332**(7556), 1463 (2006)

22. Grzymala-Busse, J.W., Grzymala-Busse, W.J.: Handling missing attribute values. In: Maimon, O., Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook, pp. 33–51. Springer, US, Boston, MA (2010)
23. Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al.: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. Jama **316**(22), 2402–2410 (2016)
24. Gwet, K.: Handbook of inter-rater reliability. STATAXIS Publishing Company (2001)
25. Haouari, B., Amor, N.B., Elouedi, Z., Mellouli, K.: Naïve possibilistic network classifiers. Fuzzy Sets Syst. **160**(22), 3224–3238 (2009)
26. Hathaway, R.J., Bezdek, J.C.: Fuzzy c-means clustering of incomplete data. IEEE Trans. Syst. Man Cybernet. **31**(5), 735–744 (2001)
27. Hayes, S.: Terminal digit preference occurs in pathology reporting irrespective of patient management implication. J. Clin. Pathol. **61**(9), 1071–1072 (2008)
28. Hemkens, L.G., Contopoulos-Ioannidis, D.G., Ioannidis, J.P.: Agreement of treatment effects for mortality from routinely collected data and subsequent randomized trials: meta-epidemiological survey. BMJ **352**, i493 (2016)
29. Hüllermeier, E.: Possibilistic instance-based learning. Artif. Intell. **148**(1–2), 335–383 (2003)
30. Hüllermeier, E.: Fuzzy sets in machine learning and data mining. Appl. Soft Comput. **11**(2), 1493–1505 (2011)
31. Hüllermeier, E.: Does machine learning need fuzzy logic? Fuzzy Sets Syst. **281**, 292–299 (2015)
32. Jha, S., Topol, E.J.: Adapting to artificial intelligence: radiologists and pathologists as information specialists. JAMA **316**(22), 2353–2354 (2016)
33. Katz, J.: The silent world of doctor and patient. JHU Press (2002)
34. Krippendorff, K.: Content analysis: an introduction to its methodology. Sage (2012)
35. Lodwick, W.A.: Fundamentals of interval analysis and linkages to fuzzy set theory, pp. 55–79. Wiley (2008)
36. Maravalle, M., Ricca, F., Simeone, B., Spinelli, V.: Carpal tunnel syndrome automatic classification: electromyography vs. ultrasound imaging. TOP **23**(1), 100–123 (2015)
37. Mitchell, T.M.: Machine learning. Burr Ridge, IL: McGraw Hill **45**(37), 870–877 (1997)
38. Obermeyer, Z., Emanuel, E.J.: Predicting the future big data, machine learning, and clinical medicine. New Engl. J. Med. **375**(13), 1216 (2016)
39. Parasuraman, R., Manzey, D.H.: Complacency and bias in human use of automation: an attentional integration. Hum. Factors J. Hum. Factors Ergon. Soc. **52**(3), 381–410 (2010)
40. Parsons, S.: Qualitative Approaches for Reasoning Under Uncertainty. The MIT Press, Cambridge, Massachussets (2001)
41. Paxton, C., Niculescu-Mizil, A., Saria, S.: Developing predictive models using electronic medical records: challenges and pitfalls. In: AMIA Annual Symposium Proceedings. vol. 2013, p. 1109. American Medical Informatics Association (2013)
42. Pivert, O., Prade, H.: A certainty-based model for uncertain databases. IEEE Trans. Fuzzy Syst. **23**(4), 1181–1196 (2015)
43. Prevention, C., et al.: For disease control, ICD-9-CM official guidelines for coding and reporting. Technical Report Centers for Medicare & Medicaid Services, Atlanta, GA, USA (2011)
44. Reiser, S.J., Anbar, M.: The Machine at the Bedside: strategies for using technology in patient care. Cambridge University Press (1984)
45. Reiser, S.J.: The clinical record in medicine Part 2: Reforming content and purpose. Ann. Intern. Med. **114**(11), 980–985 (1991)
46. Ruamviboonsuk, P., Teerasuwanajak, K., Tiensuwan, M., Yuttitham, K., for Diabetic Retinopathy Study Group, T.S., et al.: Interobserver agreement in the interpretation of single-field digital fundus images for diabetic retinopathy screening. Ophthalmology **113**(5), 826–832 (2006)
47. Shafiq, A., Arnold, S.V., Gosch, K., Kureshi, F., Breeding, T., Jones, P.G., Beltrame, J., Spertus, J.A.: Patient and physician discordance in reporting symptoms of angina among stable coronary artery disease patients: Insights from the angina prevalence and provider evaluation of angina relief (appear) study. Am. Heart J. **175**, 94–100 (2016)

48. Shortliffe, E.H., Buchanan, B.G.: A model of inexact reasoning in medicine. Math. Biosci. **23**(3–4), 351–379 (1975)
49. Simpkin, A.L., Schwartzstein, R.M.: Tolerating uncertainty the next medical revolution? New Engl. J. Med. **375**(18), 1713–1715 (2016)
50. Spodick, D.H., Bishop, R.L.: Computer treason: intraobserver variability of an electrocardiographic computer system. Am. J. Cardiol. **80**(1), 102–103 (1997)
51. Svensson, C.M., Hubler, R., Figge, M.T.: Automated classification of circulating tumor cells and the impact of interobsever variability on classifier training and performance. J. Immunol. Res. **2015** (2015)
52. Timmermans, S., Berg, M.: The Gold Standard: the challenge of evidence-based medicine and standardization in health care. Temple University Press (2010)
53. Tsumoto, S.: Medical diagnosis: rough set view. In: Thriving Rough Sets, pp. 139–156. Springer (2017)
54. van der Lei, J., et al.: Use and abuse of computer-stored medical records. Methods Archive **30**, 79–80 (1991)
55. Van Driest, S.L., Wells, Q.S., Stallings, S., Bush, W.S., Gordon, A., Nickerson, D.A., Kim, J.H., Crosslin, D.R., Jarvik, G.P., Carrell, D.S., et al.: Association of arrhythmia-related genetic variants with phenotypes documented in electronic medical records. Jama **315**(1), 47–57 (2016)
56. Veress, B., Gadaleanu, V., Nennesmo, I., Wikström, B.: The reliability of autopsy diagnostics: inter-observer variation between pathologists, a preliminary report. Int. J. Qual Health Care **5**(4), 333–337 (1993)
57. Vetterlein, T., Mandl, H., Adlassnig, K.P.: Fuzzy arden syntax: a fuzzy programming language for medicine. Artif. Intell. Med. **49**(1), 1–10 (2010)
58. Wang, Y.T., Tadarati, M., Wolfson, Y., Bressler, S.B., Bressler, N.M.: Comparison of prevalence of diabetic macular edema based on monocular fundus photography vs optical coherence tomography. JAMA Ophthalmol. **134**(2), 222–228 (2016)
59. Wong, T.Y., Bressler, N.M.: Artificial intelligence with deep learning technology looks into diabetic retinopathy screening. JAMA **316**(22), 2366–2367 (2016)