AI & SOCIETY (2020) 35:811–818
https://doi.org/10.1007/s00146-020-00945-9

ORIGINAL ARTICLE

Concordance as evidence in the Watson for Oncology decision-support system

Aaro Tupasela[1] · Ezio Di Nucci[2]

Received: 20 September 2019 / Accepted: 15 January 2020 / Published online: 1 February 2020

Abstract

Machine learning platforms have emerged as a new promissory technology that some argue will revolutionize work practices across a broad range of professions, including medical care. During the past few years, IBM has been testing its Watson for Oncology platform at several oncology departments around the world. Published reports, news stories, as well as our own empirical research show that in some cases, the levels of concordance over recommended treatment protocols between the platform and human oncologists have been quite low. Other studies supported by IBM claim concordance rates as high as 96%. We use the Watson for Oncology case to examine the practice of using concordance levels between tumor boards and a machine learning decision-support system as a form of evidence. We address a challenge related to the epistemic authority between oncologists on tumor boards and the Watson Oncology platform by arguing that the use of concordance levels as a form of evidence of quality or trustworthiness is problematic. Although the platform provides links to the literature from which it draws its conclusion, it obfuscates the scoring criteria that it uses to value some studies over others. In other words, the platform "black boxes" the values that are coded into its scoring system.

Keywords Artificial intelligence · Decision support · Machine learning · Oncology · Watson for Oncology · IBM · Clinical trials

1 Introduction: decision-support systems in healthcare

During the past several years, IBM has been developing, among others, the Watson for Oncology platform (WFO), which is an artificial intelligence cognitive computing system (see IBM 2018). Such systems are more generally called medical decision-support systems. These systems are designed to support doctors in making decisions on which treatment option is best suited for their patients based on the latest medical evidence that is available (He et al. 2019; Char et al. 2018). The system relies on natural language processing and machine learning to provide treatment recommendations.

Machine learning platforms are not new and have been operating for years within search engines, such as Google, as well as financial markets and many other everyday services (Carlson 2018; Mittelstadt 2016; Sharon 2016; Buchanan 2015). Algorithmic decision-making is becoming increasingly omnipresent in our everyday lives (Zarsky 2015). According to its proponents, one of the benefits of algorithmic decision-making is that decisions become more objective (cf. Lepri et al. 2018; D'Agostino and Durante 2018). Machine learning-based medical decision-support systems can help to identify new research findings that physicians may not have time to find out for themselves among the rapidly expanding medical literature, as well as free up time for patients themselves. Machine learning platforms can also help to speed up the identification of treatment options, help to reduce errors, provide cost-efficiency, help provide standardized care, as well as support oncologists through uncertainty and risk. Some applications have even received FDA approval for diagnosis, such as the IDx-DR device, which is used to diagnose diabetic retinopathy (Healio 2018).

Within the medical field, machine learning platforms are expected to revolutionize the way in which we study and

✉ Aaro Tupasela
aaro.tupasela@helsinki.fi

[1] Faculty of Social Science, University of Helsinki, Unioninkatu 35, 00014 Helsinki, Finland

[2] Centre for Medical Science and Technology Studies, University of Copenhagen, Copenhagen, Denmark

diagnose disease (Hinton 2018; Thrall et al. 2018; Syeda-Mahmood 2018). Domingos (2015, xi) has suggested that machine learning is different from other computational approaches in that machine learning platforms are able to "figure it out on their own, by making inferences from data." The treatment options offered by the WFO platform have been developed between IBM and Oncologists at Memorial Sloan Kettering Cancer Center in New York, which is a world-leading center for cancer treatment. Numerous efforts are underway globally to test its concordance with tumor board decisions (cf. Xu et al. 2019; Kim et al. 2019a, b; Liu et al. 2018). Concordance refers to the level to which the treatment options offered by the platform agree with the treatment options that are chosen by the oncologists at a certain hospital. It is also the metric IBM uses to advertise its platform on its website (cf. Somashekhar 2017).

Our paper is based on observations of a pilot project conducted between IBM and the Department of Oncology at Rigshospitalet in 2017, which is Denmark's leading hospital. We use the Danish experience and compare it with other pilot projects IBM has conducted around the world and that IBM itself highlights on its own Watson for Oncology webpage as success stories. We critically examine the practice of using concordance levels between tumor boards and the WFO decision-support system as a form of evidence, although we recognize the potential of the WFO platform to aid oncologists in their everyday work. We address a challenge related to the epistemic authority between oncologists on tumor boards and the Watson Oncology platform by arguing that the use of concordance levels as a form of evidence is problematic since it does not address the fundamental metric of outcome of specific treatment options for different patients. Furthermore, the WFO platform does not make visible the scoring criteria it uses to rank results from cancer clinical trials. Although WFO supports its recommendations with the literature that makes the treatment claims, we see this as masking a more fundamental problem regarding evidence: how does the WFO platform score or weight different clinical studies with regard to quality and local applicability of results? We, therefore, criticize the assumption that all data and studies that algorithms draw on are equal in quality or value (cf. Jaton 2017). This tension raises the substantive question of how to evaluate the quality of data algorithms draw evidence from and how should machine learning decision-support systems be evaluated in general.

Studies suggest that cancer trials contain significant levels of bias. Ledford (2019), for example, argues that ethnicity is a major concern for geneticists with regard to studying cancer across populations, but is often overlooked in Western clinical trials for cancer. A recent study on cancer drugs approved by the European Medicines Agency (EMA) argued that almost all studies between 2014 and 2016 were at risk of bias (Naci et al. 2019). Others have suggested that trials conducted by the pharmaceutical industry, as opposed to academic investigators alone, differ vastly with regard to their goals, whereby studies conducted by the pharmaceutical industry may not have the best interests of the patient as a concern, but rather focus on drug approval (Piccart et al. 2007). Furthermore, Devasenapathy et al. (2009) suggest that clinical trials conducted in developing countries may suffer from inadequate research infrastructure, as well as expertise to conduct robust studies. The value judgements regarding which studies ought to be weighted and for what populations are at the heart of the problem of concordance as a form of evidence in the WFO platform since WFO does not provide scoring information regarding which studies are deemed more reliable and how it accounts for bias. As such, inexperienced oncologists may lack the training and understanding of how WFO operates to critically examine the evidence that WFO provides.

As machine learning platforms begin to proliferate, we suggest that several questions emerge within the medical field. First, how should the recommendations made by WFO be evaluated in relation to physician expertise of treatment options on a global scale? Second, what can be said about the multiple levels of concordance between different tumor boards around the world in relation to Watson Oncology? Third, are there better alternatives for evaluating the effectiveness of machine learning platforms for cancer treatment other than concordance?

The pursuit of validation through global concordance levels is further complicated by discussions surrounding value-flexibility in the development of machine learning platforms (Hodgkin 2016). McDougall (2018), for example, has suggested that many artificial intelligence systems that are developed to support medical decision-making may mask fixed and covert value judgements, which erode the important role of shared decision-making in medical treatment (for a debate on this see also Di Nucci 2019; McDougall 2019). McDougall suggests that in the medical field, for example, patient perspectives are rarely considered in developing treatment option recommendations, which suggests a reintroduction of medical paternalism to patient treatment practices. Although platforms such as Watson may provide exciting new opportunities to help oncologists make decisions about possible treatment options at a global level, there is a risk that such platforms also introduce values and practices, which are not locally shared by physicians and patients alike. Following Mckinlay's assertion that "If we expect computational systems to provide us with something as complex as explanation or evidence, it seems appropriate that any epistemic assumptions we bring to bear upon such systems be subject to careful and rigorous analysis" (Mckinlay 2017, 463). The use of concordance as a form of evidence represents a type of "algorithmic culture" (Striphas 2015), which we consider problematic since it lacks reliable and comparable reference or metric through which patient outcomes can be evaluated.

2 Methods

This paper is based on 14 semi-structured qualitative interviews conducted by AUTHOR in 2017 and 2018 in Denmark. The interviews were conducted with policy-makers, hospital administrators, IBM representatives, as well as doctors who were involved in piloting or developing a broad range of machine learning platforms in their respective fields. Written informed consent was sought from the interviewees before the interview. The interviews were transcribed and thematically analyzed for themes regarding the testing use and possible implementation of decision-support systems in medical care. In addition to interviews, we have collected and analyzed scientific publications, reports, media articles, as well as online documents published by IBM related to its oncology platform. Some of the countries where WFO has been piloted include India, China and South Korea. Our analysis has focused on the way in which the notion of concordance is used as a form of evidence to generate credentials through agreement between oncologists and the machine learn ing platform.

3 Watson for Oncology

The WFO platform seeks to address an issue that some have perceived to be a problem within medical treatment. According to Somashekhar (2017) the system "processes structured and unstructured data from medical literature, treatment guidelines, medical records, imaging, lab and pathology reports, and the expertise of memorial Sloan Kettering experts to formulate therapeutic recommendations." The algorithm then suggests what treatment options the physician should use in treating the patient. The programming for the algorithm accounts for numerous different data sources that can be weighted differently. These data sources include different bio-markers derived from the patient (sex, age, type of tumor, etc.), published findings in journals, national treatment standards, etc. The system then comes up with a score regarding which treatment options are best suited for the patient.

This process might mask, however, bias and statistical errors, which are built into the system, as well as the way in which results have been reported from other pilot studies. In a recent blog post Coeira (2018), for example, provides a critical analysis of a study published by Somashekhar et al. (2018)—a study in which oncologists in an Indian hospital report high concordance rates with Watson for Oncology. In the critical review, Coeira points out numerous limitations of the study. These limitations included how the WFO study drew on a sample population which excluded many patients, failing to describe well enough what data were entered into the system and what came out of it, why concordance rates were calculated twice allowing for oncologists to change their preferred treatment options, as well as not sufficiently recognizing that the outcome of the study reflects a process and not a clinical outcome. As Coeira notes:

> "So far we have got to the point where the papers' results are not that WFO had a 93% concordance with human experts, but rather that, when humans from a single institution read cancer cases from that institution, and extract data specifically in the way that WFO needs it, and also when a certain group of breast cancers are excluded, then concordance is 93%. That is quite a list of caveats already" Coeira 2018).

In addition to the statistical criticism laid out by Coeira, the WFO raises a number of other concerns regarding the usefulness of using concordance as a form of evidence. First, the platform has thus far been developed on US-based treatment guidelines (Kim et al. 2019a, b) and does not take into account differences in incidence or risk of certain kinds of cancers in different populations (Liu et al. 2018). Yet, different countries, such as Denmark, base their treatments on their own national guidelines that have been established by local expert panels. These nationally based expert panels also discuss, debate, and interpret the latest findings in international studies. Consequently, expert panels form value judgements on different studies based on their perceptions regarding the quality of the evidence and rigor of the methods used. Using the Sloan Kettering treatment options, the IBM programmers and developers are creating a de facto 'ground truth' (Jaton 2017) against which other treatment options are measured (Kim et al. 2019a, b). This 'ground truth' is a type of bias that is introduced into the program (Gianfrancesco et al. 2018), since it is not compared or validated against other treatment standards. In and of itself, this approach may not be problematic, so long as the treatment standards used in the US can be proven the best treatment standards in the world. This variance between guidelines reflects value judgements that medical experts make in each local setting in relation to the preferred treatment options and evidence, which they see as trustworthy and relevant. IBM could address this bias using population-based studies, as well as long-term prospective studies, for example, or explaining the logic behind their scoring system. The Nordic countries, for example, are some of the only countries in the world, which maintain cancer registries for the whole population (Tupasela et al. 2020). These data can be linked to personal health and treatment data, which provide a far more robust data set, albeit limited in terms of its applicability to other populations (Pukkala et al. 2018, 2009).

Second, the global market for cancer treatments is not unitary. Not all medications and treatments are available or have been approved in a uniform manner in different markets. Therefore, there may be treatment options in the US that are not available in different parts of the world. This may also work the other way, there may be treatment options available elsewhere that have not been approved in the US yet.

Third, in reading the pilot studies that have been run in different countries, it becomes evident that IBM regularly releases new versions of its software. Liu et al. (2018), for example, report that they have used version 17.1 of the WFO software; while, Kim et al. (2019a, b) have tested version 18.9. In trying to compare these results, it is difficult to ascertain how IBM has changed its program between different versions. The WFO website does not provide information about version history, which essentially generates a black box regarding the scoring system. Many of the published studies do not even report the version number, making it even more difficult to compare different iterations of the WFO algorithm.

Fourth, in reading the published reports regarding WFO, one must note how different oncology teams use different sets of data points, which they entered into WFO. The Danish oncologists reported that they provided 17 data points, whereas Kim et al. (2019a, b) provided 13. Although not specifically a major problem, we see this as another situation in which opacity is introduced into the scoring system.

The opportunities and benefits that Watson may offer lie in its ability to cover a great deal of literature and offer a service for countries where there may not be enough resources for local oncology panels to evaluate the latest findings. Its weakness, however, lies in the black boxing of epistemic value judgements regarding which studies can be considered of good quality and how they are weighted in the algorithmic decision process.

3.1 Treatment guidelines and concordance

When visiting the Watson Oncology website (https://www.ibm.com/watson/health/oncology-and-genomics/oncology/), one is presented with several case studies from around the world where various tumor boards discuss the level of concordance (agreement) that their local protocols have with that of the platform. According to IBM "Recent studies continue to demonstrate that Watson for Oncology 'agrees' with physicians around the world in the vast majority of cases—so experts can focus on what they do best—deliver care" (IBM 2018). The presented studies have been given as conference presentations of "early experiences" with working with the platform. None of the studies represents scientific articles that have been peer-reviewed, although numerous such studies have been published to date (Kim et al. 2019a, b; Liu et al. 2018; Xu et al. 2019). Hospitals which have piloted the platform and whose preliminary findings are presented on the website include institutions in India, Mexico, and Thailand. The findings that are reported in the scientific presentations include statements such as the following: "In a double-blinded study, the doctors at Manipal Hospitals found that Watson was concordant with the tumor board recommendations in 90% of breast cancer cases" (Somashekhar et al. 2016; see also Somashekhar et al. 2018).

When the platform was piloted in Copenhagen, the results of the treatment protocols were not as good as many had hoped. In comparing the suggested treatment protocols of the 31 virtual patients to the recommended treatment options in use at the hospital, only about 30% matched the current best practice in Denmark, about 30% were somewhat similar, and 30% completely different from the treatment options that the oncologists would themselves use to treat patients. These findings are similar to those reported by Kim et al. (2019a, b) and Liu et al. (2018). In further discussions between the oncologists and the IBM oncology team, two main reasons were highlighted to account for this low level of concordance. First, the platform was coded to US-based treatment guidelines, which meant that it did not account for nationally derived guidelines or preferences for treatment options. It is unclear, thus far, whether the levels of concordance would be higher if this change was to be applied to the platform. Second, the algorithm was programmed in such a way as to emphasize cancer treatment studies, which were based in the US. Liu et al. (2018) highlight how such preference can generate problems when applying the recommendations to non-US populations. They note that there is a "large difference between the EGDR gene mutation phenotype of lung cancer in China compared with that in Western countries" (Liu et al. 2018, 6). According to the Danish informants, the use of US-based studies was problematic given that the Danish oncologists considered many of them to be biased. As one doctor noted in an interview:

> We only tested it in 31 patients and well it was some of a disappointment for us. The advice from Watson was seriously flawed in 1/3 and 1/3 it was okay, but we did not agree completely and in 1/3 we had complete agreement on the advice. Watson for oncology only gives advice on which treatment should you offer the patient. Nothing else. So we have discussed a lot why the reason for that and probably the local adaptation that is missing from Watson. It was tested and developed in the United States. The practices in United States seems to be so different. So even if looking at evidence which is part of the Watson technology that the American doctor, the board they use to score the level of evidence in the literature are very biased towards USA studies. So studies we completely dis-

> card in Europe, they look at as important studies in United States (Interview with a physician 2017).

According to the Danish doctors, the US guidelines did not match the Danish best practice guidelines, and the US-based studies were not considered as good as the ones conducted in Denmark. The Danish doctors argued that the lack of quality in the US studies was due to their lack of systematic nation-wide cancer registries and healthcare records; whereas, the Danish and other Nordic country studies were considered far more robust and representative in relation to validity (Pukkala et al. 2018, 2009). As such, the Danish oncologists placed a great deal more value on studies and clinical trials conducted in Denmark, or the Nordic countries, where they believed the results were more valid for the local national populations than those conducted in the US (cf. Timmermans and Berg 2003).

Other hospitals, besides those in Denmark, have also had less success in finding high concordance rates with Watson. In South Korea, for example, several pilot studies at different hospitals found relatively low concordance rates (~40%) between the oncologists and Watson for different types of cancer (Choi et al. 2019; Lee et al. 2018; Ross and Swetlitz 2017). The outcome of the pilot studies resulted in little interest by those hospitals to adopt the platform into their everyday care practices (Choi 2018).

It is interesting to note, however, that in the interviews which were conducted in Denmark, doctors did mention that such a platform may be useful for oncology departments or oncologists who did not have the level of expertise and facilities which were available in countries such as Denmark. As Coeira (2018), however, points out "The evidence in this study doesn't yet support such a conclusion. We have some data on concordance, but no data on how concordance affects human decisions, and no data on how changed decisions affects patient outcomes." This conclusion is also supported by a number of studies, which have tested the WFO platform (Kim et al. 2019a, b; Liu et al. 2018). Interestingly, however, Liu et al. (2018) note that the WFO platform has been introduced to more than 70 medical institutions throughout China despite the challenges that have been identified thus far regarding its reliability. As such, Watson may provide, in the future, benefits to countries where expert panels are not available to develop and evaluate national standards for care; however, without further evidence it is impossible to say whether the resource would be of benefit to the patients. It is another question, however, if such countries would also have access to the types of treatment options that Watson offers in the first place as well. This possibility, however, remains a lesser issue in relation to the question of concordance as a type of evidence for a medical platform. In the following section, we will discuss this problem in more detail.

3.2 Discussion: concordance as a form of evidence in medical decision-making

As a form of evidence that IBM uses to market its product, concordance is challenging from a medical decision-making perspective. We see concordance as a form of evidence problematic for a number of reasons. Let us consider these problems through the following hypothetical scenarios regarding concordance and non-concordance.

First, if Watson is piloted with a group of oncologists and there is a high level of concordance between the two, what does this high level of concordance suggest? In the best-case scenario, both the Watson platform and the oncologists have chosen treatment options, which give the best survival rate outcomes possible given the state of knowledge at a given time. Watson's role in this decision-making process then is more to confirm what the oncologists already knew and as such does not provide any new information that can be used to treat patients.

In a second scenario, there is still a high level of concordance between the platform and oncologists. In this scenario, however, both the platform and the oncologists choose treatment options that result in poor survival outcomes, or at least outcomes that are worse than other available treatment options, but neither is "aware" of this. The fact that there is concordance between Watson and the oncologists does not provide evidence that the treatment options are the best ones; only that the two forms of expertise agree. In this scenario, the oncologists are given a false sense of security because of the high concordance level between the two. Furthermore, the developers of the platform are led to believe that because there is concordance with the oncologists, the platform must be getting it "right", thereby exacerbating the mistake.

In a third scenario, Watson provides the best possible treatment options, but there is a low level of concordance with the oncologists. In this scenario, the oncologists chooses sub-optimal treatment options while the platform does what it is hoped to do: suggest the best possible treatment options available. In this scenario, the oncologists can choose one of two options; either discard their own expertise or discard the expertise of Watson. By following the options suggested by Watson, they end up saving more lives. By discarding Watson's suggestions, they end up harming patients with sub-optimal treatment options. One could argue that Watson was developed with this scenario in mind, being able to provide new options that the oncologists did not know about, but which end up saving more lives than the treatment options that the doctors are currently using. For the developers of Watson, however, it is unclear what non-concordance implies; is there some evidence that it is not aware of, are the calculations wrong, or are there different values at stake in calculating treatment preferences? And for the doctors there is the crucial epistemological question of whether they have access to the relevant information, data and literature to be able to evaluate Watson's suggestion and learn from it.

In a fourth scenario, Watson provides poor treatment options, while the options used by the doctors are optimal to saving lives. Here again, the doctors need to choose between following their own expertise or the options that the platform provides. By following their own expertise, patients are saved, by choosing what Watson suggests they end up harming patients. This scenario is similar to that of the third one in that the physicians need to evaluate the validity of their own expertise in relation to that which the platform provides. Again, for the developers, there is a challenge in understanding why there is a low level of concordance between the platform and the oncologists.

Although these scenarios are contrived and do not necessarily represent real-world situations, they none-the-less point to an inherent problem that the use of concordance as a form of evidence has in the marketing of medical decision-making platforms. Concordance does not measure patient outcomes regarding different treatment options. As mentioned, non-concordance does not imply that Watson is wrong and may indeed lead to oncologists learning something new. As a form of evidence, however, concordance says little about the outcomes of chosen treatment protocols, which would be of greater value for determining which options to choose. Nor does concordance give us any information regarding how and why physicians would or would not change their treatment decisions based on Watson's suggestions.

Watson's merits lie in the fact that it can cover a great deal of literature in a very short time. One could argue that in the long run, Watson could be a useful tool in comparing the long-term outcomes of different treatment protocols to see which ones may be more effective in treating different types of cancers. Such an undertaking would require, however, a better understanding of how data for cancer studies are collected and analyzed in different context, such as, for example, Denmark vs the US. There are also other important values, which the platform should account for, such as the preferences of patients as to what treatment options are preferable in different situations, which Watson does not account for (McDougall 2018).

4 Conclusion

The use of concordance levels as the only form of evidence in marketing is a problematic metric for validity of treatment options. The challenge regarding the WFO platform is that oncologists do not know what value judgements IBM codes into their algorithm regarding the scoring of different cancer studies. We have shown that clinical trials for cancer drugs contain many biases, which IBM does not make transparent in the platform. IBM could easily address this problem by making the scoring and ranking criteria more transparent. This would align the WFO platform with recent policy calls for transparency, diversity, non-discrimination as well as fairness in developing AI (European Union 2019). Still, it should be nonetheless emphasized that the opacity issue and the algorithmic bias issue are related but independent from each other, in such a way that more transparency does not necessarily guarantee less bias.

Although Watson is able to perform many important tasks relating to the collection and screening of a large amount of medical literature regarding cancer treatment options, it is still unable to evaluate the role of value judgements made by different oncology guidelines and clinical trials regarding the best treatment options. These value judgements include what studies are considered as valid in evaluating treatment options, which treatment options are made available through insurance and national health insurance schemes, as well as whether patient perspectives are included in selecting preferred treatment options. In revisiting our original questions, the use of concordance as a form of evidence has the following problems:

First, the Watson for Oncology platform is unable to provide criteria or explanation for the differences in expert judgment in different contexts. Consequently, it is difficult to evaluate why specific treatment options are ranked higher than others. Second, low concordance rates do not necessarily mean that patients in a given hospital are receiving poor treatment. All that can be said about it are that protocol choices differ between Watson and a particular set of oncologists. As we show with our scenarios, differences in concordance can have multiple explanations. Third, a better alternative would be to draw evidence from high-quality and validated outcomes. Such studies would include population studies, as well as prospective cohort studies. Although, Watson does this partially by considering the outcomes of studies, the platform needs to take more into account differences in value judgements and quality criteria that different oncologists and national boards for treatment options take into consideration. As we note, the quality and reliability of cancer clinical trials vary greatly, with many studies containing biases (Copur 2019). Despite providing evidence through literature, WFO does not reveal its scoring system and how it accounts for possible bias in studies. Such functionality could be added to Watson and provide interesting new insight for oncologists around the world regarding differences in treatment options that oncologists choose and for what reason. As IBM is able to pilot and test its platform around the world in different contexts, it is also able to amass a large amount of data regarding different iterations of its algorithm; that is to say different versions that may prioritize one form of evidence over another. Making this information available would make evaluation of the platform easier and more transparent. Fourth, our example highlights how all data sets and studies are not of equal value. Instead, researchers, as well as oncologists, use numerous different quality criteria to evaluate the validity of oncology studies. With the WFO platform, oncologists do not know what the scoring criteria are regarding which studies are more relevant to others.

Following McDougall (2018), considerations for bias are not accounted for enough in developing machine learning platforms for medical decision-support systems (see also Ledford 2019; Devasenapathy et al. 2009). Many of the people listed in the early findings' reports have also disclosed funding and ownership relations to IBM. For such platforms to gain broader acceptance and validity, the platform needs to provide more transparency regarding the scoring criteria it uses, as well as the changes that different versions bring with it. Programmers and developers also need to consider value-flexibility in developing systems, which prioritize treatment options.

The platform does provide some interesting new opportunities, however. Through its piloting projects around the world, IBM could develop a repository of data regarding the different treatment guidelines and applications used in different countries, as well as the values that lie behind such treatment guidelines. As such, IBM is in a unique position to gather data that could hypothetically be used later to compare different guidelines. To evaluate and develop evidence as to the best treatment guidelines, however, IBM would need to trace the outcomes of the patients over many years to see which guidelines generated the best survival and re-occurrence outcomes among the patient populations. It is imperative that the suggestions made by Watson can be critically examined in light of competing treatment guidelines and data quality to ensure that the suggestions that it is making are the best possible options available globally, not just in the US and by a small set of oncologists.

Finally, we would like to conclude by addressing a broader problem which the Watson for Oncology platform appears to point towards. Many of the machine learning platforms that are developed to support medical decision-making seem to reproduce expert knowledge, which is already known. This, one could argue, is not machine learning, but rather a form of automation (Wajcman 2017). This raises the question of which set of values ought to be relied on as guiding decision-making? This suggests that decision-support systems within medicine might always have to rely on locally rooted value judgements, regardless of whether IBM is able to offer high concordance rates between the oncologists and its platform. We suggest that further research on how value judgements are made in these different contexts will help to provide valuable insight into the quality of care that patients receive in different regions of the world.

Acknowledgements Open access funding provided by University of Helsinki including Helsinki University Central Hospital.

References

Buchanan M (2015) Trading at the speed of light. Nature 518:161–163

Carlson M (2018) Automating judgment? Algorithmic judgment, news knowledge, and journalistic professionalism. New Media Soc 20(5):1755–1772. https://doi.org/10.1177/1461444817706684

Char DS, Shah NH, Magnus D (2018) Implementing machine learning in health care—addressing ethical changes. N Engl J Med 378:981–983

Choi MH (2018) Major Hospitals in S. Korea not very interested in Watson. 2 March 28, 2018. http://www.businesskorea.co.kr/news/articleView.html?idxno=21308. Accessed 14 May 2019

Choi YI, Chung J, Kim KO et al (2019) Concordance rate between Clinicians and Watson for Oncology among patients with advanced gastric cancer: early, real-world experience in Korea. Can J Gastroenterol Hepatol. https://doi.org/10.1155/2019/8072928 **(Article ID 8072928)**

Coeira E (2018) Journal review: Watson for Oncology in Breast cancer. The Guide to health informatics 3rd edn. https://coiera.com/2018/03/09/journal-review-watson-for-oncology-in-breast-cancer/. Accessed 26 Aug 2019

Copur MS (2019) State of cancer research around the globe. Oncology 33(5):181–185

D'Agostino M, Durante M (2018) Introduction: the governance of algorithms. Philos Technol 31(4):499–505

Devasenapathy N, Singh K, Prabhakaran D (2009) Conduct of clinical trials in developing countries: a perspective. Curr Opin Cardiol 24(4):295–300

Di Nucci E (2019) Should we be afraid of medical AI? J Med Ethics 45:556–558

Domingos P (2015) The master algorithm. Penguin Books, London

Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G (2018) Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med 178(11):1544–1547

He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K (2019) The practical implementation of artificial intelligence technologies in medicine. Nat Med 316:30–36

Healio (2018) FDA allows marketing of IDx-DR artificial intelligence device for diabetic retinopathy detection. www.healio.com. Accessed 13 May 2018

Hinton G (2018) Deep learning—a technology with the potential to transform health care. JAMA 320(11):1101–1102. https://doi.org/10.1001/jama.2018.11100

Hodgkin PK (2016) The computer may be assessing you now, but who decides its values. BMJ 355:6169

IBM (2018) Product vignette: IBM Watson for Oncology. https://www.ibm.com/watson/health/oncology-and-genomics/oncology/. Accessed 28 May 2018

Jaton F (2017) We get the algorithms of our ground truths: designing referential databases in digital image processing. Soc Stud Sci 47(6):811–840. https://doi.org/10.1177/0306312717730428

Kim M, Kim BH, Kim JM, Kim EH, Kim K, Pak K, Lee BJ (2019a) Concordance in postsurgical radioactive iodine therapy recommendations between Watson for Oncology and clinical practice in patients with differentiated thyroid carcinoma. Cancer 125:2803–2809

Kim D, Kim YY, Lee JH, Chung YS, Choi S, Kang JM, Chun YS (2019b) A comparative study of Watson for Oncology and tumor boards in breast cancer treatment. Korean J Clin Oncol 15(1):3–6

Ledford H (2019) Cancer geneticists tackle ethnic bias in studies. Nature 568(7751):154–155

Lee et al (2018) Assessing concordance with Watson for Oncology, a cognitive computing decision support system for colon cancer treatment in Korea. JCO Clin Cancer Inform 2(2):1–8

Lepri B, Oliver N, Letouzé E, Pentland A, Vinck P (2018) Fair, transparent, and accountable algorithmic decision-making processes. Philos Technol 31(4):611–627

Liu C, Liu X, Wu F, Xie M, Feng Y, Hu C (2018) Using artificial intelligence (Watson for Oncology) for treatment recommendations amongst Chinese patients with lung cancer: feasibility study. J Med Internet Res 20(9):e11087

McDougall RJ (2018) Computer knows best? The need for value-flexibility in medical AI. J Med Ethics. https://doi.org/10.1136/medethics-2018-105118

McDougall RJ (2019) No we shouldn't be afraid of medical AI; it involves risks and opportunities. J Med Ethics 45:559

Mckinlay ST (2017) Evidence, explanation and predictive data modelling. Philos Technol 30(4):461–473

Mittelstadt BDM (2016) Auditing for transparency in content personalization systems. Int J Commun 10:4991–5002

Naci H, Davis C, Savović J, Higgins JP, Sterne J, Gyawali B et al (2019) Design characteristics, risk of bias, and reporting of randomised controlled trials supporting approvals of cancer drugs by European Medicines Agency, 2014–16: cross sectional analysis. BMJ 366:l5221

Piccart M, Goldhirsch A, Wood W, Pritchard K, Baselga J, Reaby L, Coates A (2007) Keeping faith with trial volunteers. Nature 446(7132):137–138

Pukkala E, Martinsen JI, Lynge E, Gunnarsdottir HK, Sparén P, Tryggvadottir L, Kjaerheim K (2009) Occupation and cancer-follow-up of 15 million people in five Nordic countries. Acta Oncol 48(5):646–790

Pukkala E, Engholm G, Højsgaard Schmidt LK, Storm H, Khan S, Lambe M, Malila N (2018) Nordic cancer registries—an overview of their procedures and data comparability. Acta Oncol 57(4):440–455

Ross C, Swetlitz (2017) IBM pitched its Watson supercomputer as a revolution in cancer care. It's nowhere close. https://www.statnews.com/2017/09/05/watson-ibm-cancer/. Accessed 12 May 2019

Sharon T (2016) The Googlization of health research: From disruptive innovation to disruptive ethics. Pers Med 13(6):563–574

Somashekhar SP et al (2017) Early experiences with IBM Watson for Oncology (WFO) cognitive computing system for lung and colorectal cancer. J Clin Oncol 35(15_suppl):8527 **(San Antonio Breast Cancer Symposium, December 9th, 2016)**

Somashekhar SP et al (2016) Validation study to assess performance of IBM cognitive computing system Watson for oncology with Manipal multidisciplinary tumour board for 1000 consecutive cases: an Indian experience. Ann Oncol. https://doi.org/10.1093/annonc/mdw601.002

Somashekhar SP et al (2018) Watson for Oncology and breast cancer treatment recommendations: agreement with an expert multidisciplinary tumor board. Ann Oncol 29(2):418–423. https://doi.org/10.1093/annonc/mdx781

Striphas T (2015) Algorithmic culture. Eur J Cult Stud 18(4–5):395–412

Syeda-Mahmood T (2018) Role of big data and machine learning in diagnostic decision support in radiology. J Am Coll Radiol 15(3PB):569–576

Thrall JH et al (2018) Artificial intelligence and machine learning in radiology: opportunities, challenges, pitfalls, and criteria for success. J Am Coll Radiol 15(3PB):504–508

Timmermans S, Berg M (2003) The gold standard. The challenge of evidence-based medicine and standardization in health care. Temple University Press, Philadelphia

Tupasela A, Snell K, Tarkkala H (2020) The Nordic data imaginary. Big Data Soc. https://doi.org/10.1177/2053951720907107

Union European (2019) Trustworthy AI—joining efforts for strategic leadership and societal prosperity. European Commission, Brussels

Wajcman J (2017) Automation: is it really different this time? Br J Sociol 68(1):119–127

Xu F, Sepúlveda MJ, Jiang Z, Wang H, Li J, Yin Y, Song Y (2019) Artificial intelligence treatment decision support for complex breast cancer among oncologists with varying expertise. JCO Clin Cancer Inform 3:1–15

Zarsky T (2015) The trouble with algorithmic decisions. Sci Technol Hum Values 41(1):118–132