pHealth 2022
B. Blobel et al. (Eds.)

doi:10.3233/SHTI220965

The Causal Plausibility Decision in Healthcare

Vije Kumar RAJPUT[ac], Mette Kjer KALTOFT[b], Jack DOWIE[ab,1]
[a] *London School of Hygiene and Tropical Medicine*
[b] *University of Southern Denmark*
[c] *Stonydelph Medical Centre, Tamworth, UK*

Abstract. The explosion of interest in exploiting machine learning techniques in healthcare has brought the issue of inferring causation from observational data to centre stage. In our work in supporting the health decisions of the individual person/patient-as-person at the point of care, we cannot avoid making decisions about which options are to be included or excluded in a decision support tool. Should the researcher's routine injunction to use their findings 'with caution', because of methodological limitations, lead to inclusion or exclusion? The task is one of deciding, first on causal plausibility, and then on causality. Like all decisions these are both sensitive to error preferences (trade-offs). We engage selectively with the Artificial Intelligence (AI) literature on the causality challenge and on the closely associated issue of the 'explainability' now demanded of 'black box' AI. Our commitment to embracing 'lifestyle' as well as 'medical' options for the individual person, leads us to highlight the key issue as that of who is to make the preference-sensitive decisions on causal plausibility and causality.

Keywords. Artificial Intelligence (AI), black box, machine learning, unsupervised, causality, explainability, causability, clinical decision support, individualisation, personalization

Preamble

In 2019, Caroline Kramer and colleagues from the University of Toronto Mt Sinai Diabetes Center published the results of a meta-analysis of an observational data set [1]. A meta-analysis is a relatively 'shallow' form of machine learning, not seen as warranting the 'black box' characterization of the 'deep learning' forms of Artificial Intelligence (AI) now flourishing. In her acknowledgements, Caroline thanked Romeo Kramer, her miniature Schnauzer dog, for inspiring the meta-analysis, which examined the association of dog ownership with all-cause mortality, with and without prior cardiovascular disease.

"Ten studies were included yielding data from 3 837 005 participants (530 515 events; mean follow-up 10.1 years). Dog ownership was associated with a 24% risk reduction for all-cause mortality as compared to non-ownership (relative risk, 0.76; 95% CI, 0.67–0.86) with 6 studies demonstrating significant reduction in the risk of death. Notably, in individuals with prior coronary events, living in a home with a dog was associated with an even more pronounced risk reduction for all-cause mortality (relative

[1] Corresponding author, Jack Dowie, LSHTM, 15-17 Tavistock Place, London, UK WC1H 9SH; Email: jack.dowie@lshtm.ac.uk

risk, 0.35; 95% CI, 0.17–0.69; I2 , 0%)... Dog ownership is associated with lower risk of death over the long term, which is possibly driven by a reduction in cardiovascular mortality…. A possible limitation was that the analyses were not adjusted for confounders" [1, p. 1].

As one might expect, especially in the light of this last limitation, a critical response followed. Eventually, the editor of the journal was moved to publish a set of commentaries, making valid, albeit entirely predictable, comments [2]. A response from the authors [3] included some sensitivity and confounder-adjusted analyses in support of their conclusion, but importantly (from the point of view of the present paper) they also introduced a biological rationale for the meta-analytic findings.

"… Specifically, dog ownership has been shown to increase owner's physical activity, lower blood pressure, reduce physiological indices in response to stress, and improve mental health among others positive effects beyond the cardiovascular system. In addition, the existence of a special human-dog bond was demonstrated in an experiment published in *Science*[7] in which an oxytocin-gaze positive loop was observed between dogs and their owners that was similar to the hormonal response observed when mothers interact with their infants. In conclusion, while we recognize that the current evidence consisting of mostly observational data cannot establish any causal association between dog ownership and survival, we disagree that dog ownership is solely a marker of increased physical activity. Given the growing body of evidence, we speculate that such positive interspecies interaction promotes greater adherence to a healthy lifestyle, improves cardiophysiological indices, improves overall well-being, and promotes the activation of a hormonal loop associated with positive emotions that are likely beneficial to human health [3, p. 815]."

We do not need to go further into the various and valid methodological concerns in this case, because we take the multiple uncertainties they give rise to as the given normal in healthcare decision making. (This applies even when high quality interventional studies have been undertaken.) For the researcher, the ultimate conclusion is typically that more research is needed. But, also typically, and perhaps a tad gratuitously, they will add the injunction that the present results should be *used* 'with caution'.

1. Introduction: Individual Decision Support

Our work is exclusively on decisional use, so we do not have the luxury of retaining the scientific purity of the researcher, remaining agnostic as to whether a link is causal or not. Having read Kramer, and being fully cognizant of all the issues raised in the ensuing debate, we are faced by an individual, with or without a prior coronary event. What should we do about dog ownership in our decision support tool? Should it be included as an option, along with, as we do for medication options, a performance rating on the all-cause mortality criterion. Or not? Of course, beyond this specific question lies the general one that has nothing to do with any particular intervention, such as owning a dog. That general and unavoidable decision as to whether a *causal* link between interventions and outcomes exists - has been given new and heightened exposure by the arrival of 'deep learning' AI. So, while this paper is prompted by the recent excitement about the potential for AI to improve healthcare, the issue addressed is eternal.

Our interest is in the improvement of all *individual* health decisions, taken inside or outside a healthcare system, and anywhere within it, but we use primary care as our exemplar 'inside' setting, usually without great loss of generality. We also approach the topic in the specific context of providing decision support to both professional and person (including patient-as-person), the latter being regarded as the *decision owner.*

We need to establish our position on individual decision support right away. (We interpret individual decision support to embrace clinical decision support). We define the best possible decision support as that which both maximally *individualizes* and maximally *personalizes.* Maximal individualization is achieved by obtaining and inputting the best available estimates for the option *performance ratings,* i.e. how well each of the adoptable causes of actions, including doing nothing, performs on each of the harms and benefits that matter to the specific person (i.e. their criteria). Maximal personalization is achieved by eliciting and inputting the specific person's criteria, along with their criterion *importance weightings* for each harm and benefit (i.e. their preferences). Given that they are fundamentally distinct ontologically, these two sets of inputs (performance ratings and importance weightings) must be produced independently, only subsequently being integrated into the necessary overall assessment (*'score')* for each option. Individual decision support tools which achieve these objectives facilitate, perhaps virtually ensure, care that meets the legal 'reasonable patient' standard for informed consent [4].

2. Artificial Intelligence (AI)

Our focus on improving decisions rather than adding to knowledge has important implications for our response to the AI literature. It means immediate attention is required to the distinction between teleological and causal explanations.

The explanation for professionals and persons undertaking decision making - and deciding whether or not to engage with an individual decision support tool in the process - is teleological, not causal. As agents, neither is *caused* to engage in decision making, but do so in pursuit of a goal (Greek *telos* purpose, end), being assumed to be seeking the best means of achieving it. However, in pursuit of that goal (*effect)* and, under inevitable uncertainty, they will be interested in determining what means may *cause* it and with what *probability*. It is in facilitating the best estimates of causal probability, and only in such facilitation, that we believe AI can make a major potential contribution to improved individual healthcare and health. This contribution is limited by failure or reluctance to accept that it can be manifested - 'add value*'* - only in the teleological setting of decision making, not simply by 'adding knowledge*'* to a research depository.

There are two noteworthy features of the seriously reflective literature on AI in healthcare, which we will take to be exemplified in the survey article by Eric Topol [5]. One is its almost total absorption with AI applications in relation to medical techniques (such as scans) and medical treatments (such as drugs). There is relatively little AI interest in 'lifestyle' databases, partly, perhaps mainly, a knock-on effect of the failure to invest in collecting 'hard big data' about lifestyles, as compared with devices and drugs. Second, there is almost exclusive concern with the quality of the AI-based information that would potentially be passed to decision makers, and a taken-for-granted assumption that the acceptable quality threshold for passing is to be set by researchers, not decision makers/owners. It is never suggested or contemplated that the appropriate quality threshold should be determined in the context of the individual decision. We demur, and will be returning to this point at length, since it is crucial for our work in developing *individual* decision support tools. Tools that must include non-medical options, such as pet ownership, where we believe the issue of *whose* causality thresholds are to be applied is likely to become more and more prominent and contentious.

Finally, by way of background, we take the view that all data production, collection, exploration, and processing is conducted on the basis of some implicit or explicit *model*, i.e. a constructed 'small world'. All empirical results and interpretations are produced within a 'small world' model and reflect underlying conceptual constructions and methodological preferences. The single most important motivation for all data activity in this small world is to *somehow* improve *something*. In the case of healthcare, including our decision support, that something is a construction of 'health'. We find all-cause mortality to be an uncontroversial component of most health constructs and so focus on this for the present purpose, albeit it is only one of the multiple criteria relevant in most health decisions.

3. The demand for 'trustable AI'

There is now a vast literature on the enhanced techniques for the centuries-old activity of data exploration, developed and implemented under the umbrella term AI. As 'machine learning' (ML) from observational data becomes more and more technically sophisticated - and incidentally profitable - concern with what is going on in 'deep learning' programs has increased in parallel. It has now reached the point where the orthodox position has become a demand for *'trustable* AI'. The 2022 MIE conference had this in its title and theme and it is useful to note for future use, how the issue was presented as one of control, with the problems - and solutions - located solely in the AI system.

"… new machine learning methods…promise to improve the accuracy of diagnosis and screening, support clinical care, and assist various public health interventions such as disease surveillance, outbreak response, and health system management. Naturally, as these new AI systems emerge, concerns arise concerning the level of control that should be conceded when reviewing the pace at which AI methods are introduced. One concern in particular arises from the fact that these new AI systems often work as "black boxes", and are unable to explain their results. Explainability is critical, however, to respond to patient and practitioner narrative exchanges, and the fact that practitioners, who are responsible for their decisions, cannot easily follow the proposals of AI systems that they disagree with is also a problem [6, p. v]."

Before addressing the central issue of what is to constitute 'explainability', it is important to reaffirm our position that AI can only improve decision support through improving the *individualization* (of the option performance *ratings*). To do so, AI exercises must, of course, have addressed the well-known problems of confounding that arise in non-interventional sources; but since they invariably seek to do so, to the extent possible, we do not need to spend time on this. However, in line with what was said earlier, the use of AI results must avoid trespassing into *personalization* and/or *decision making* conceptualized as the final *integration* of individualized ratings and personalized weightings. Such trespassing will inevitably occur when the data exploited in the AI exercise reflect the *preference* judgements embedded in the decision that generated each datum in the data set. This is likely to be particularly important when the data reflect patterns of practice, based on the trade-offs among outcomes *assumed* to be appropriate by guideline bodies. Here, unfortunately, we can do little more than call attention to this neglected, non-epidemiological confounder, perhaps more serious than many of the recognized epidemiological and biophysical confounders of which AI researchers are well aware.

The call for 'trustable AI' is a call for an *explanation* of what is going on inside the AI 'black box'. It is not necessary to delve deeply into the many possible constructions of the terms 'explanation', and closely associated ones, such as 'explainability' and 'explainee' exhaustively documented by Carvalho and colleagues [7]. We need only to make clear that any conclusions emerging from these conceptual exercises, can only inform, not determine, the appropriate requirements (standards) for 'explainability', in a specific context and for a specific decision owner. This follows from the fact that setting these standards are *decisions* and therefore logically involve value judgments and preferences, above all and inevitably, judgments on the trade-offs made among the main types of possible error. This logic is perfectly captured in Ken Hammond's brilliant subtitle to his *Human Judgment and Social Policy*: *Irreducible Uncertainty, Inevitable Error, Unavoidable Injustice* [8].

The orthodox demand, as in the MIE quote above, is for 'explanations' that go beyond technical *procedural* accounts, such as 'we conducted a Bayesian network analysis'. It is maintained, correctly, that such accounts will not be 'understandable' by many, perhaps almost all explainees and lead to their characterization as 'black boxes'. Such *procedural* accounts will also not provide the *declarative* information - about *what* has been found to be causing *what*, and to what extent - which is desired/required by professionals and patients. To anticipate, we will argue that thinking these two demands can be met within a single 'explanation' is based on a serious conceptual confusion.

In response to these demands a number of 'Explainable AI' (eXAI) programs have been developed to delve into deep machine learning programs [9]. Programs such as SHAP are able to extract the feature/category/attribute importance (coefficients) detected by a deep learning program and present these *declarative results* in visually effective ways - undoubtedly making them more 'understandable' and 'interpretable' to a much wider audience of explainees. One of the many data sets used in teaching about AI techniques and eXAI programs is one relating to the predictors of survival on the Titanic. It contains eight passenger categories and in most exercises the prediction coefficients (in SHAP, the 'Shapley values') for sex and class of passage emerge as the most important, with being female and travelling first class 'explaining' increasing survival to the extent and in the direction indicated by these values.

Two things about such eXAI programs require emphasizing. Firstly, their impressive accomplishment in presenting these 'causal coefficients' must not distract from the fundamental point that they do not in any way actually *explain the process* by which these findings were arrived at. They simply present a summary plot of these findings in a visually communicative and hence more interpretable and useful way (including useful for us in our pursuit of performance ratings). The supply of the (demanded) declarative information about *what* the program has found to cause *what*, and to *what extent*, is widely confused with an explanation of *how* the program found these. That latter explanation can only be a *procedural* one, of the sort which most people find hard to understand and therefore characterize as a 'black box'. Secondly, eXAI programs can only produce findings in relation to the features/categories /attributes entered into the analysis, and not those which were not; in other words, findings in relation to those features/categories/attributes which met some threshold for causal plausibility and were included. We do not know whether data relating to the pets accompanying Titanic passengers does not exist, or was regarded as lacking in causal plausibility, and was therefore not processed in the ML application or the SHAP interrogation of it.

In some cases causality may be ‘beyond reasonable doubt’ to all (reasonable) observers. This will be particularly the case with biophysical causes (e.g. trauma, poison). At the other extreme, the impossibility of causality may be beyond reasonable doubt, except in some ‘placebo’ sense (e.g. snake oil). (For the record, we do not question the evidence on the effectiveness of placebos; however, we see the cause of the placebo effect as the prescriber, not the prescribed). In between these extremes, there are the vast range of *possible* causes which have to be assessed for whether they meet some standard for causal plausibility, or ‘causability’ in Holzinger’s term [10].

In the case of Judea Pearl’s campaign for ‘strong AI’, in *The Book of Why* [11], this assessment determines whether the category/feature/attribute is entered into the causal diagram that he insists must be drawn *prior* to any AI exploitation of a data set. This causal model will therefore determine (as ‘supervisor’ of the ML program) the scope of the processed set, often confining it to a subset of the fuller set that could be explored unsupervised (e.g., excluding information about the pets of Titanic passengers, if this was readily available.) And so the findings of ML programs will, by definition, be restricted to those attributes selected *ex ante*. Pearl therefore rules out the ‘weak AI’ in which causal plausibility is assessed *ex post* in order to determine whether the outputs of an unsupervised exercise are to be accorded causality.

We can find no reason to disassociate ourselves from this reviewer’s summary of The Book of Why.

“One of the book's central claims, re-asserted again and again, is that data on their own are "dumb," and cannot be causally interpreted until one draws a diagram, representing a set of assumptions about the reality behind the data. In Pearl's view, the assumptions in a causal diagram live in some separate, a priori realm, completely distinct from "data." He never tells us where the diagrams are supposed to come from, if not from empirical observations about the world; the book contains scattered references to "background knowledge" or "common sense", but these never coalesce into a general statement about what sort of information is allowed to inform our choice of diagram. We are only told that whatever this information is, it must not be "data."

So, in Pearl's version of causal inference, you must first choose a diagram before you see the observational data at all. You are not allowed to change this diagram once you see the data, since diagrams do not come from data… Although he never says as much… the entire subject of Pearl's book [is] no more and no less than estimating the coefficients for a pre-specified, fixed diagram… we are adrift in [a] strange world where one makes certain causal assumptions (encoded in a diagram) for the sake of assessing others, never sure whether any given arrow is an inviolable assumption or a testable hypothesis, or what makes the difference [12].”

But our inference from this, and weaker versions of this argument, should not be misinterpreted. Our fundamental point is not that we should avoid intermediate binary categorization of a category as ‘plausibly causal/not plausibly causal’ for a target outcome (Titanic survival, All-Cause Mortality). Nor that we should avoid final categorization of it as ‘causal/not causal’. In decision making we cannot. At some point, ‘drawing a causal diagram’ - constructing a causal model - is unavoidable and should be undertaken very seriously. The basic assumption and rationale of this paper is that all such models of causation must be seen as constructs produced by *preference-sensitive decisions* of human beings. Under unavoidable uncertainty, assigning causality involves setting a threshold on a scalar variable of causal plausibility. In any such threshold-setting exercise, the possibility of type 1 and type 2 errors arises, with a false positive resulting from assigning causality to a category which (according to a *constructed* gold standard/'ground truth') has none, a false negative resulting from denying causality to a category which is causal. Any threshold setting can therefore only be done on the basis of error trade-off *preferences* - how serious are the consequences of a false positive relative to those of a false negative?

These preferences can have no basis in 'science', 'knowledge' or 'expertise'. The type 1/type 2 trade-off appropriate when seeking the *truth*, in scientific research, is completely different from that appropriate when seeking the *best option* in the individual decision, *now*, *under uncertainty*, *in practice*. This trade-off will depend on all the circumstances of the case and should reflect the individual decision owner's preferences, not those of the professional - nor the profession [13] - and certainly not that of any expert group of researchers.

In the provision of individual decision support in healthcare we see the consequences of human decisions on inclusion or exclusion from the intermediate 'plausibly causal' set and hence the final causal model as potentially substantial. Few advocates of 'trustable AI' see that the elephant in the room is *who*se preferences are to be entered in setting the decision threshold. The (unwritten) Book of Who is more important than *The Book of Why*.

4. The Data Sets

It may help if we present a simplified filter taxonomy for the data sets that might be used in machine learning (ML) to contribute to healthcare.

> "1 Data for a category (characteristic/attribute) which *is not* produced/collected/ 'cleaned', but only because of excessive practical difficulty and/or expense.
>
> 2 Data for a category (characteristic/attribute) which *could be* produced/collected/ 'cleaned', without much practical difficulty and at acceptable expense, *but is not*
>
> A. because of its universally perceived irrelevance to any healthcare evaluation
> B. because its collection is prohibited by some legal, political, or professional authority on ground/s unconnected with its causal plausibility
> C. because it lacks causal plausibility in the eyes of an 'expert' group and would therefore be excluded *ex ant*e from a 'supervised' ML exercise – so there is no point
>
> 3 Data which can be collected without practical difficulty, is indeed collected, but is excluded (*ex ante*) from a 'supervised' ML exercise as lacking causal plausibility.
>
> 4 Data which can be collected without practical difficulty, is indeed collected, but is subsequently dismissed (*ex post*) from a 'supervised' ML exercise as lacking causal plausibility.

> 5 Data which remain - those for a category (characteristic/attribute) which are deemed to reach some causal plausibility threshold."

An empirical illustration of filtering is provided by a Danish research group that used AI techniques to identify the predictors of 'acute critical illness' in the electronic health record of all residents of four mixed rural and urban municipalities (18 years or older in 2012–2017) [14]. The potentially vast data set was

"… limited to include 27 laboratory parameters and six vital signs. The parameters were selected by trained specialists in emergency medicine (medical doctors) with the sole purpose of simplifying the model to enable a better discussion of the model explanations of its predictions [of 'acute critical illness]. While a deeper model with more parameters might lead to better performance, it would also have made the discussions between clinicians and software engineers difficult. Therefore, the scope of this article is not to obtain the best performance at all costs but to demonstrate how clinical tasks can be supported by a fully explainable deep learning approach [14, p. 5]."

Their final comments (below) are in line with the group signing up to the orthodox assumption that simply fails to question whether clinicians have the competencies required to understand what they need to understand in order to understand.

"Clinicians must be able to understand the underlying reasoning of AI models so they can trust the predictions and be able to identify individual cases in which an AI model potentially gives incorrect predictions. Consequently, a useful explanation involves both the ability to account for the relevant parts in an AI model leading to a prediction, but also the ability to present this relevance in a way that supports the clinicians causal understanding in a comprehensible way. An explanation that is too hard to perceive and comprehend will most likely not have any practical effect [14, p. 1]."

This hubristic assumption is dangerous in itself, apart from its restriction to 'medical' causes and explanations. Moreover, like most of the eXAI literature, the Danish study falls prey to the fallacy that *displaying* the causal findings of an AI program constitutes an explanation of how it arrived at these causal findings. It does not. The latter requires the deep *procedural explanation* that has already been proclaimed to be too difficult to understand!

With its comprehensive unique personal identifier system, Denmark is the perfect setting for an unsupervised ML exercise on a massive population-wide database. Among the large number of non-zero coefficients that might emerge for predicting all-cause mortality is that for dog ownership. We will never know, if we don't undertake it.

5. Judging Probable Cause

If humans are to judge causal plausibility and causality in the absence of an interventional experiment, as we routinely have to, what do we know about how we do this, as experts or non-experts of all kinds? In an early seminal paper of 1986, Einhorn and Hogarth proposed an ABC by which we judge probable cause - using Alternatives, Background, and Cues:

"… judgments of probable cause are affected by a causal background, probabilistic cues-to-causality, and a discounting process for dealing with specific alternatives… these general ideas can be summarised by a perceptual analogy in which causal candidates are differences-in-a-background (figures are seen against ground), good explanations arise from internally consistent patterns of cues (good figures form a gestalt), and good explanations have few plausible alternatives (as do good figures) [15].”

The go-to paper on causality in the psychological literature is now the Sloman and Lagnado 2015 review of *Causality in Thought* [16]. Given our entirely practical interest in developing individual decision support tools to be used in decision making practice, we confine ourselves to this central question they ask:

“Can human causal inference be captured by relations of probabilistic dependency, or does it draw on richer forms of representation?”

In answering this *descriptive* question, they reviewed research in reasoning, decision making, various forms of judgement and attribution, and concluded:

“We endorse causal Bayesian networks as the best normative framework and as a productive guide to theory building [about causal thinking]. However, it is incomplete as an account of causal thinking. On the basis of a range of experimental work, we identify three hallmarks of causal reasoning—the role of mechanism, narrative, and mental simulation—all of which go beyond mere probabilistic knowledge… causal inference is not merely a way of representing and updating probabilities; it is not merely Bayesian inference. Human causal inference involves the construction of narratives that unfold over time and determine the focus of attention, narratives that reflect knowledge of the specific mechanisms that drive effects…the causal Bayes nets framework has not offered a silver bullet that answers all questions about human thought. We have seen that it is not always clear whether a choice should be modeled as an intervention [16, pp. 223, 236, 240].”

We fully acknowledge all the descriptive points made by Sloman and Lagnado, though point out no one has ever suggested that the causal Bayes nets framework answers *all questions about human thought*. However, it *may* answer or contribute significantly to answering the question in which we are interested. As established earlier, we decided that our *practical* individual decision support tools, designed for use, not theorizing, should incorporate and display the best available estimate of the degree to which (in many cases, probability) treatment options will produce (cause) a specified effect (criterion). So, in contrast to Sloman and Lagnado we find ourselves satisfied with ‘mere probabilistic knowledge’. We do not even begin to consider how ‘mechanism’, ‘narrative’ and ‘mental simulation’ could be incorporated, *along with probabilistic dependencies*, in our individual decision support tool. And, for the avoidance of all doubt, this individual decision support tool is designed to support choice, modelled as a set of possible interventions to address the question ‘what should we do, now?’ We would be happy to trial our type of individual decision support tool against one in which ‘mechanism, narrative and mental simulation’ are embodied, on a level playing field. Examples of our individual decision support tools can be found at https://easybest.org.uk, with introductions in [17, 18].

6. Why is CI Treated Differently from AI?

The blackness of a black box is in the eye of the beholder; the transparency of an AI program is in the mind of the ‘explainee’. What the output of a black box is telling the ‘explainee’, with great clarity if supplemented by a program like SHAP, is not an explanation of how it came to arrive at that output. Holzinger's example of the application of his System Causability Scale [19] (see Appendix), developed in the context of his call for a *discipline* of causability [10] illustrates this very well. A doctor is asked to assess the Framingham Risk Tool on this 10 item 5 level Likert Scale. This doctor scores it at .86. However, he has been asked nothing about the procedure by which the Framingham tool arrives at its cardiac risk score.

The clinical encounter is a classic illustration of another 'black box' in operation. However, a clinician is never expected to have to explain their 'clinical judgement', a term for which we substitute Clinical Intelligence (CI) to emphasize the contrast with AI. Clinical Judgement-as-CI is, by definition, differentiated from any explicit reasoning process or analytical account and invariably constructed as an holistic intuitive process. Its assumed existence is exemplified in the NICE statement that precedes every guidance it issues [https://www.nice.org.uk/guidance]. The clinician *must* exercise their judgement in every case, and

"When exercising their judgement, professionals and practitioners are expected to take this guideline fully into account, alongside the individual needs, preferences and values of their patients or the people using their service."

In contrast to the calls for the' black box' of AI machine learning process to be opened up and examined for their 'explainability', 'interpretability' and 'understandability', clinicians are never faced by similar calls in relation to their CI process; the process which enables them, inter alia, to 'take this guideline fully into account, alongside the individual needs, preferences and values of their patients'. The call is wisely not made, because it could not be answered, except as a *theory* of the process, constructed *ex post.* Geoff Hinton, head of Google Brain, and renowned for his work with Artificial Neural Networks, is very clear:

"People can't explain how they work, for most of the things they do... People have no idea how they do that. If you ask them to explain their decision, you are forcing them to make up a story [20, p. 3]."

Process tracing by psychologists and 'expert systems' researchers long ago established that expert clinicians are unable to provide a procedural account of their decisions as IF-THEN production rules, or indeed within any other analytical framework. The source of holistic CI (which we agree is undoubtedly possessed, though to varying degrees, by clinicians) is generally assumed to reside in the intuitive skill of 'pattern recognition' acquired through experience. Here, our point is that, *as a procedural explanation,* 'pattern recognition' is unlikely to satisfy anyone other than possibly other clinicians - or those who 'trust' them to the extent that no demand is made on them for causal plausibility or 'explainability'. Since pattern recognition is an important basis for machine learning, whether or not it becomes an acceptable 'explanation' of findings in the world of AI will be interesting to observe.

7. The Lost Opportunities

The Marshall-Warren discovery of the effectiveness of antibiotics in treating stomach ulcers is perhaps the most famous - notorious - example of how professional resistance to accepting even *plausible causality* can hinder healthcare progress [21]. It parallels the well-documented reluctance of the clinical community to accept the outcome of the clinical vs actuarial debate on *predictability* [22].

There are manifold examples of 'opportunistic' findings in healthcare research leading to confirmatory research and improved treatment. Liraglutide for bile acid diarrhea is a recent one [23]. But such findings need not be opportunistic. Unsupervised deep learning programs set to work on massive databases should be seen as abductive generators of plausible causes, to be followed by unbiased *independent* assignment of causal plausibility *in a defined task*. At the moment the 'raw' individual data collected in health research of all kinds is both defined and processed in accordance with a discipline-based study protocol.

In the current research set-up/industry these protocols invariably specify the aim of the study as being to evaluate/assess/describe, 'quantitatively', 'qualitatively' or both, some way or ways of improving something or some things. The something could be 10 year fracture risk for people with chronic obstructive pulmonary disease, or 'shared decision making processes in primary care'. It doesn't matter what it is. It could and should be exploring *all* the possible causes within the data set in relation to generic outcomes, all-cause mortality and all-cause morbidity being the main candidates. The opportunity is lost through the targeted framing of the research and silo-based 'ring-fencing' in terms of what data is collected, how it is processed, and what happens to it subsequently. These restrictions mean that gigabytes of potentially exploitable data are simply being left unexploited, either silently or with various justifications. Justifications which include that they are 'commercial in confidence', but more often ones that reflect the non-financial material interests of researchers and their corporative organizations.

We are deliberately using 'silo', in order to stress that the compartmentalization of research has massive downsides, as well as the upsides which lead to its existence. These downsides are swept under the carpet by unspoken/taken-for-granted agreement among all disciplines that such compartmentalization ('specialization') is in their individual and collective material interests, in terms of growth in funding, jobs, publications, and kudos.

Characterizing generic unsupervised exercises pejoratively, as 'trawling' or 'fishing expeditions', is in line with the material interests of silo-based research groups.

What can be done? Individuals are perhaps the most likely agents of change. In line with the spirit of mydata.org (https://www.mydata.org) they could insist that their data, even when it is being collected for a cause-specific purpose, be inserted into unsupervised machine learning programs targeted at generic outcomes. Individuals could be encouraged to agree to participate in research only on the condition that their contributions (their *raw* data, quantitative or qualitative) are made available in pseudo-/anonymized form for 'OPEN' unsupervised AI learning, independent of its silo origin, within x years of its collection. At an institutional level, research protocols could be approved by Ethics Committees only when it is shown how data inputs provided *pro bono* by voluntary participants will be fully exploited in relation to *all* health benefits and harms.

8. Postamble

Should Caroline have acquired her dog and should she be proud of owning Romeo, a miniature Schnauzer, as she states. The answers, arrived at by simple logical human reasoning, is that it all depends on (a) the things that matter to her - her multiple criteria - and their relative importance - the weights she attaches to 'health' and 'non-health' criteria such as all-cause mortality and feeling proud; and on (b) how having a dog and 'being proud' effects (performs on) each of these criteria. We cannot support her in this decision if some group of people have pre-emptively decided that there will be little or no information available on (b), because possessing a dog has not met *their* standards for causal plausibility and causality. Such pre-emption should not be in their power, since the necessary error trade-offs should be the preferences of Caroline, the decision owner.

The core fallacy, perpetrated in evidence-based medicine and evidence-based guidelines, is that the error trade-off appropriate in setting a threshold for admittance to the status ‘true’, has some legitimate connection with the error trade-off appropriate in a decision about what should be done now, under whatever uncertainties exist. While our focus is on the individual, this actually applies in group level decisions, such as in public health, where the issue of causation within epidemiology has built on the pioneering work of Bradford Hill, but notably without attention to the error trade-offs necessary in decision making [24]. The WHO statement on AI is similarly free of attention to the preference-sensitivity of decisions [25]. Whose preferences are the key issue in all decision contexts. The cognitive sources of flaws in decision making both group and individual, have been brilliantly summarized and illustrated in NOISE - A flaw in human judgment [26]. Unfortunately, addressing them will not necessarily have any impact on the motivational and political biases embedded in the power structure of the status quo. It is these that will determine the preference basis of decisions - including whether or not to address them We should be wary of allowing the medical research tail to wag the health decision maker dog [27].

References

[1] Kramer CK, Mehmood S, Suen RS. Dog Ownership and Survival: A Systematic Review and Meta-Analysis. *Circ Cardiovasc Qual Outcomes*. 2019 Oct;12(10):e005554. doi: 10.1161/CIRCOUTCOMES.119.005554. Epub 2019 Oct 8. PMID: 31592726.

[2] Nallamothu BK. Hard Science: Editor’s Perspective *Circ Cardiovasc Qual Outcomes*. 2020 Oct;13(10):e007359. DOI:10.1161/CIRCOUTCOMES.120.007359 Epub 2020 Oct 20. PMID: 33079587.

[3] Kramer CK, Mehmood S, Suen RS. Response by Kramer et al to Letters Regarding Article, "Dog Ownership and Survival: A Systematic Review and Meta-Analysis". *Circ Cardiovasc Qual Outcomes.* 2020 Oct;13(10):e006388. doi: 10.1161/CIRCOUTCOMES.119.006388. Epub 2020 Oct 20. PMID: 33079587.

[4] Chan SW, Tulloch E, Cooper ES, Smith A, Wojcik W, Norman JE. Montgomery and informed consent: where are we now? BMJ. 2017 May 12;357:j2224. doi: 10.1136/bmj.j2224. PMID: 28500035.

[5] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med.* 2019 Jan;25(1):44-56. doi: 10.1038/s41591-018-0300-7. Epub 2019 Jan 7. PMID: 30617339.

[6] Séroussi B, Weber P, Dhombre F, Grouin C, Liebe J-D, Pelayo S, et al (eds) Challenges of Trustable AI and Added-Value on Health. Proceedings of MIE 2022 IOS Press BV, Amsterdam.

[7] Carvalho DV, Pereira EM, Cardoso JS. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics* 2019, 8, 832; doi:10.3390/electronics8080832.

[8] Hammond KR. *Human Judgment and Social Policy*: *Irreducible Uncertainty, Inevitable Error, Unavoidable Injustice* 1996 Oxford University Press.

[9] ElShawi R, Sherif Y, Al-Mallah M, Sakr S. Interpretability in healthcare: A comparative study of local machine learning interpretability techniques. *Computational Intelligence* 2021 37 (4) 1633-50 doi.org/10.1111/coin.12410.

[10] Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip Rev Data Min Knowl Discov*. 2019 Jul-Aug;9(4):e1312. doi: 10.1002/widm.1312. Epub 2019 Apr 2. PMID: 32089788; PMCID: PMC7017860.

[11] Pearl J, Mackenzie D. *The Book of Why: The New Science of Cause and Effect* 2018 Basic Books.

[12] nostalgebraist. Review of *The Book of Why* https://www.goodreads.com/book/show/36204378-the-book-of-why.

[13] Dowie J, Wildman M. Choosing the surgical mortality threshold for high risk patients with stage Ia non-small cell lung cancer: insights from decision analysis. *Thorax*. 2002 Jan;57(1):7-10. doi: 10.1136/thorax.57.1.7. PMID: 11809982; PMCID: PMC1746179.

[14] Lauritsen SM, Kristensen M, Olsen MV, Larsen MS, Lauritsen KM, Jørgensen MJ, Lange J, Thiesson B. Explainable artificial intelligence model to predict acute critical illness from electronic health records. *Nat Commun*. 2020 Jul 31;11(1):3852. doi: 10.1038/s41467-020-17431-x. PMID: 32737308; PMCID: PMC7395744.

[15] Einhorn H, Hogarth R. Judging Probable Cause. *Psych Bull.* 1986 99 (1): 3-19.

[16] Sloman SA, Lagnado D. Causality in thought. *Annu Rev Psychol.* 2015 Jan 3;66:223-47. doi: 10.1146/annurev-psych-010814-015135. Epub 2014 Jul 21. PMID: 25061673.

[17] Dowie J, Rajput V, Kaltoft MK. A Generic Rapid Evaluation Support Tool (GREST) for Clinical and Commissioning Decisions. Stud Health Technol Inform. 2019 Aug 21;264:576-580. doi: 10.3233/SHTI190288. PMID: 31437989.

[18] Rajput VK, Kaltoft MK, Dowie J. A Multi-Criterial Support Tool for the Multimorbidity Decision in General Practice. Stud Health Technol Inform. 2019;261:205-210. PMID: 31156117.

[19] Holzinger A, Carrington A, Müller H. Measuring the Quality of Explanations: The System Causability Scale (SCS): Comparing Human and Machine Explanations. *Kunstliche Intell* (Oldenbourg). 2020;34(2):193-198. doi: 10.1007/s13218-020-00636-z. Epub 2020 Jan 21. PMID: 32549653; PMCID: PMC7271052.

[20] Jones H. Geoff Hinton dismissed the need for Explainable AI: 8 experts explain why he's wrong. 2018 *Forbes* Dec 20.

[21] Abbott A. Gut feeling secures medical Nobel for Australian doctors. *Nature.* 2005 Oct 6;437(7060):801. doi: 10.1038/437801a. PMID: 16208334.

[22] Dawes RM, Faust D, Meehl PE. Clinical versus actuarial judgment. Science. 1989 Mar 31;243(4899):1668-74. doi: 10.1126/science.2648573. PMID: 2648573.

[23] Kårhus ML, Brønden A, Forman JL, Haaber A, Knudsen E, Langholz E, Dragsted LO, Hansen SH, Krakauer M, Vilsbøll T, Sonne DP, Knop FK. Safety and efficacy of liraglutide versus colesevelam for the treatment of bile acid diarrhoea: a randomised, double-blind, active-comparator, non-inferiority clinical trial. *Lancet Gastroenterol Hepatol*. 2022 Jul 19:S2468-1253(22)00198-4. doi: 10.1016/S2468-1253(22)00198-4. Epub ahead of print. PMID: 35868334.

[24] Shimonovich M, Pearce A, Thomson H, Keyes K, Katikireddi SV. Assessing causality in epidemiology: revisiting Bradford Hill to incorporate developments in causal thinking. Eur J Epidemiol. 2021 Sep;36(9):873-887. doi: 10.1007/s10654-020-00703-7. Epub 2020 Dec 16. PMID: 33324996; PMCID: PMC8206235.

[25] World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. 2021 Geneva. Licence: CC BY-NC-SA 3.0 IGO.

[26] Kahneman D, Sibony O, Sunstein CR. NOISE: A Flaw in Human Judgment. 2021. London William Collins

[27] Kaltoft MK, Nielsen JB, Eiring Ø, Salkeld G, Dowie J. Without a reconceptualisation of 'evidence base', evidence-based person-centered healthcare is an oxymoron. *European Journal for Person Centered Healthcare* 2015 3 (4) 496-502**.**

Appendix: System Causability Scale (SCS) [19, P. 196]

The purpose of our SCS is to *quickly* determine whether and to what extent an explainable user interface (human–AI interface), an explanation, or an explanation process itself is suitable for the intended purpose. [Likert scale, 10 items with 5 levels from 'strongly agree' to 'strongly disagree']

1. I found that the data included all relevant known causal factors with sufficient precision and granularity.
2. I understood the explanations within the context of my work.
3. I could change the level of detail on demand.
4. I did not need support to understand the explanations.
5. I found the explanations helped me to understand causality.
6. I was able to use the explanations with my knowledge base.
7. I did not find inconsistencies between explanations.
8. I think that most people would learn to understand the explanations very quickly.
9. I did not need more references in the explanations: e.g., medical guidelines, regulations.
10. I received the explanations in a timely and efficient manner.

As an illustration, SCS was applied by a medical doctor from the Ottawa Hospital to the Framingham Risk Tool (FRT). FRT was selected as a classic example of a prediction model that is in use today. FRT estimates the risk of coronary artery disease in 10 years for a patient without diabetes mellitus or clinically evident cardiovascular disease, and uses data from the Framingham Heart Study. FRT includes the following input features: sex, age, total cholesterol, smoking, HDL (high density lipoprotein) cholesterol, systolic blood pressure and hypertension treatment. The [doctor's] ratings for the SCS score are reported in Table 1. [Total score was 86/100].