Causal Inference from Clinical Experience

Hamed Tabatabaei Ghomi and Jacob Stegenga

Forthcoming, *Philosophical Studies*

Abstract

How reliable are causal inferences in complex empirical scenarios? For example, a physician prescribes a drug to a patient, and then the patient undergoes various changes to their symptoms. They then increase their confidence that it is the drug that causes such changes. Are such inferences reliable guides to the causal relation in question, particularly when the physician can gain a large volume of such clinical experience by treating many patients? The evidence-based medicine movement says no, while some physicians and philosophers support such appeals to first-person experience. We develop a formal model and simulate causal inference based on clinical experience. We conclude that in very particular clinical scenarios such inference can be reliable, while in many other routine clinical scenarios such inferences are not reliable.

1. Introduction

Maria does not feel well, so she visits her doctor, who diagnoses a disease and prescribes a drug. Maria returns home and starts taking the drug. Time passes, and Maria's symptoms change. Maria then returns to her doctor, who determines that either Maria has improved or she has not improved, and the doctor's confidence in the drug accordingly increases or decreases. We want to know how reliable such inferences are when deployed across a series of such interactions, representing the putative knowledge that the physician might gain from their clinical experience.

Specifically, we want to know if (and if so, under what conditions) first-person clinical experience is a reliable basis for inferences about the general effectiveness of interventions. The so-called evidence-based medicine movement says that such inferences are not reliable, because of phenomena such as the placebo effect and expectation bias that influence both a patient's experience and a physician's evaluation

of that experience. Instead, says the evidence-based medicine movement, we need to test such causal relations using methods such as randomised trials that minimise the effect of such biases (Howick 2011). Historically, inferences about the effectiveness of interventions based on clinical experience have led us astray, claims the evidence-based medicine movement. This is part of the rationale for maintaining randomized trials and meta-analyses of such trials at the top of so-called evidence hierarchies, and relegating physician expertise and judgement to the bottom of such evidence hierarchies. Standard guidance from evidence-based medicine methodologists is to assess the effectiveness of interventions only with randomized trials, with explicit guidance that evidence from any other study design, including case reports, can be ignored (Blunt 2015). The view that clinical experience is an unreliable guide for inferring the general effectiveness of interventions has been widely asserted for decades (e.g. Meehl 1986; Guyatt et al. 1992; Choudhry, Fletcher, and Soumerai 2005).

On the other hand, many physicians and patients routinely make such inferences, and some physicians and philosophers have argued that appeals to first-person clinical experience can be reliable evidence for making inferences about the effects of interventions. Tonelli and Shapiro, for example, argue that clinical experience provides what they call experiential knowledge, and such knowledge is important for treatment decisions and assessing the response of treatments—they claim that expertise developed through first-person clinical experience can inform physicians how to "deploy therapeutic decisions in an optimal and individualized manner" (Tonelli and Shapiro 2020, p. 76). They also claim that "assessing the effect of an intervention is also highly dependent upon the experiential knowledge of the clinician" (p. 76). Similarly, Healy (2011) argues that physicians routinely can make reliable single-case causal inferences about the effects of interventions based on clinical experience. This view is reflected in some large surveys of physicians' attitudes to clinical experience (Dewitt et al. 2021). And Cartwright (2017) has articulated a variety of kinds of evidence that can be used to warrant single-case causal inferences.

Sometimes the appeal to causal inferences based on clinical experience is made in the context of responding to sceptical arguments about medicine; for example, Stegenga (2018) offered such a sceptical argument by appealing to small effect sizes from randomised trials and biases in those trials, and Healy (2020) responded by explicitly asserting that physicians can reliably observe the effects of drugs based on clinical experience. Healy claims that "(e)verything we have in medicine is built on professional and patient anecdotes. Every discovery of a benefit or other effects of drugs comes from this. Other evaluative techniques, and especially randomized controlled

trials (RCTs), are less accurate and less objective" (2020, p.1). Yet that view clashes with the position of evidence-based medicine, aptly summarized by Howick, who claims that clinicians are routinely led astray when assessing the effectiveness of interventions based on clinical experience "due to the natural course of illness and the placebo effect" (2011 p. 164). The aim of this paper is to offer some insight on this polarized debate, and to suggest, albeit in preliminary terms, a path through the debate which respects the concerns of both sides, while ultimately offering a partial but not full vindication of the evidence-based medicine view regarding causal inference from clinical experience.

Assessing the reliability of causal inference based on clinical experience by empirical evidence is of limited value, because precisely what is in question is the reliability of one mode of evidence (first-person experience) compared to others (randomised trials, for example). Moreover, empirical evidence about clinical experience is limited to the range of events that one can observe in actual clinical practice, which would be limited in scope for various contingent reasons about the practice one observed. In this paper, instead, we develop a formal model of causal inference from clinical experience. We then use a computer simulation to generate data based on this model, with the aim of providing insight into the reliability of causal inference from clinical experience in a range of clinical scenarios. We use simulations in line with Mayo-Wilson and Zollman (2021), as alternatives to thought experiments; thought experiments are fine, but simulations can serve a similar function while exploring a great range of counterfactual possibilities with precision. As far as we know the reliability of causal inference from clinical experience has not been evaluated using simulation, though simulations have been used to address many questions in the philosophy of science and social epistemology (for some excellent recent examples see Heesen 2018; Zollman 2015; Rubin 2022).

After Maria returns home from her doctor and begins taking the drug, there are several possible causes of changes to her symptom profile. One is whatever physiological effects the drug elicits. Another is the familiar placebo effect. Giving a clear account of what placebo is, which permits a clear distinction between the physiological effects of the drug and the placebo effects, has been a challenge for some philosophers (Grünbaum 1986, 1981; Howick 2017). For our purposes it is enough to say that placebo effects are any effects of any kind of intervention that operate through expectation effects, regardless of whether that intervention also has other physical effects. Finally, the mere passage of time, in the absence of any physiological effect or placebo effect, can involve a myriad of other causes to changes in Maria's symptom

profile—all those other causes, from her body healing itself to Maria eating a little better and sleeping a little more, we call 'the natural course of disease'.

So, there are three possible causes (or sets of causes, if you prefer) to Maria's change in symptom profile over time: the physiological effects of the intervention, the placebo effect, and the natural course of disease. The strength of each of these three causes depends on a multitude of contextual features, and particularly depends on what disease Maria has. If she has the common cold, then the natural course of disease will rapidly lead to improvement, within a week say, and even the most effective interventions will have barely any causal impact on her symptoms, as will the placebo effect. If she has depression, then the natural course of disease will likely contribute to a gradual improvement, the placebo effect will cause a strong mitigation of her symptoms, and the physiological effects of the best interventions will likely be very modest to non-existent (see Cuijpers, Stringaris, and Wolpert 2020). If she has an especially bad bacterial infection which is sensitive to antibiotic treatment, then that treatment will have a strong mitigation of symptoms, though the placebo effect and natural course of disease will have very modest effects.

For many diseases there are a range of possible interventions to choose from. For example, in the United Kingdom there are eight serotonin reuptake inhibitors available for prescription, five statins, and six fluoroquinolones – and these are each just one kind of intervention in a broader category, as for instance antibiotics include not only fluoroquinolones but also penicillins, cephalosporins, tetracyclines, and others. In our model a physician must choose from ten possible drugs. When Maria returns to the clinic after using one of these drugs, the physician's confidence in the effectiveness of that drug increases or decreases, depending on whether Maria has improved or not. Then the next patient with the same disease visits the physician, and the physician must again choose among the drugs. We repeat this for many patient visits, thereby modelling a physician's career of experience about this class of drugs. The primary question we ask is how close are physicians' estimates of the effectiveness of drugs to the stipulated effectiveness of the drugs in various clinical scenarios. We also ask how many prescriptions are required before patients improve in the different clinical scenarios.

The model we develop here is idealized in various ways, and our plan for future research is to develop the model to make it more realistic, by, for example, modelling the effect of confirmation bias and prior knowledge on the strength of placebo and natural course of improvement. Our results here, though they should be interpreted with caution, are striking: we show that if a disease is somewhat placebo-responsive or has some degree of natural course of improvement, then physicians overestimate the

effectiveness of interventions after gathering clinical experience, but when a disease has little to no placebo responsiveness and natural course of improvement, such inferences can be reliable. Thus, our results vindicate both the critics of clinical expertise in some kinds of scenarios and the proponents of clinical expertise in other kinds of scenarios.

2. Modelling Causal Inference from Clinical Experience

In our model, a patient with a particular disease visits a physician, and the physician prescribes one of k possible drugs (in the simulations here we set $k = 10$). If, after using the drug, the patient improves, this provides the physician some first-person evidence that the drug is effective. Accordingly, the probability that the physician will choose that drug in the future increases. On the other hand, if, after that drug, the patient doesn't improve, the physician gains some first-person evidence that the drug is ineffective, and so the probability of choosing that drug for subsequent patients decreases. If such first-person inferences are reliable, a seasoned physician would develop reliable judgments about the effectiveness of the drugs.

We model the physician's choice as an instance of the so-called 'multi-armed bandit problem'. Suppose you are in a casino with many slot machines, each with an unknown probability of giving you a fixed reward. In each round, you try one of the machines, and you either get a reward or not, and this provides you with some evidence from which you can infer the reward probability of each of the machines. The multi-armed bandit problem asks what the best strategy in choosing different slot machines is so that you maximise your reward over time. If you knew the reward probability of the slot machines from the outset, the best strategy would be obvious: always choose the slot machine with the highest reward probability. But you do not know these probabilities; at the beginning you know little about those probabilities and as you try out various slot machines you learn a little more about those probabilities. This adds another dimension to your decision of the slot machine you want to try. You need to balance two factors in your choice of slot machines: *exploiting* the information you have attained from your choices thus far to get rewards and *exploring* other machines so that you come up with a more accurate estimate of reward probabilities and to minimize the chance of missing a better slot machine.

The multi-armed bandit problem has many variations, applications, and solutions (Kuleshov and Precup 2014; Bouneffouf, Rish, and Aggarwal 2020). For example, it

has been used in designing clinical trials and allocating patients to treatments (Lai 1987; Villar, Bowden, and Wason 2015). Examples of earlier application of bandit problems in philosophy of science are Zollman 2007 (see also Šešelja 2022 for discussion), Holman and Bruner 2015, and Weatherall and O'Connor 2021. Reviewing these applications and solutions is beyond the scope of this paper. Here we use Thompson Sampling, a well-established and widely used Bayesian solution to the multi-armed bandit problem. Besides its computational merits as a solution to the multi-armed bandit problem, we choose this solution because we think it is a good model to represent the way that an ideal Bayesian physician would perform her clinical inferences.

Thompson Sampling

Thompson sampling is a general solution to the multi-armed bandit problem. In our specific case, we set and run it as follows. The physician starts by an initial guess about the probability of the patient improving by any of the k number of drugs. This is the physician's initial prior subjective probability and can be informed by factors such as the medical literature about the drugs, the physician's education, or other experience with the drugs. In the simulations presented here, at the beginning of each 'career', each physician starts with a uniform distribution for probability of improvement across all the drugs (this can be varied in future work). The only new source of information that subsequently informs the physician's judgement about the effectiveness of the drugs is the physician's clinical experience in using the drugs on a series of patients.

The prior distributions associated with each drug is set as a Beta distribution, which is a probability distribution defined on the interval between 0 (no chance of improvement) and 1 (certain improvement), and has two parameters α and β that determine its shape. Initially, these parameters are set to $\alpha = 1$, and $\beta = 1$ and that results in a uniform distribution on the range of 0 and 1.

$$Beta(\alpha = 1, \beta = 1)$$

Next, the physician draws one sample from the improvement probability distributions that she has associated with each of the drugs, and among the resulting hypothetical improvement probabilities, she chooses the drug with the highest hypothetical chance of causing improvement and prescribes that drug for the patient.

After trying the drug, the patient either shows improvement ($X = 1$) or does not ($X = 0$) (for simplicity, in our model improvement is binary). The physician adjusts her

prior probability that this drug is effective based on this observation. The posterior probability after the tried drug is:

$$Beta(\alpha = 1 + X, \beta = 1 + 1 - X)$$

If the patient shows improvement, the physician stops the treatment, and moves to the next patient. If the patient does not show improvement, that drug is taken out of the set of possible drugs to prescribe, and the physician repeats this process at most ten times, and then moves to the next patient. The physician gradually forms her expert opinion on the effectiveness of each of the drugs. In short, with respect to the formal aspects of inference, the physician is an ideal reasoner.

The algorithm can be summarized as follows:

1. Assume prior distributions of the probability of improvement for each of the drugs.
2. Draw a random number from each of those distributions.
3. Choose the drug that gave the highest perceived probability of improvement (highest value among the random draws in the previous step).
4. Prescribe that drug.
5. Update the probability distribution of effectiveness of that drug based on the observation of improvement or no improvement (get posterior distributions of the probability of improvement).
6. If the patient has not shown improvement, and if you have tried less than ten prescriptions for this patient, go to step 2.
7. Start treating the next patient by going to step 2.

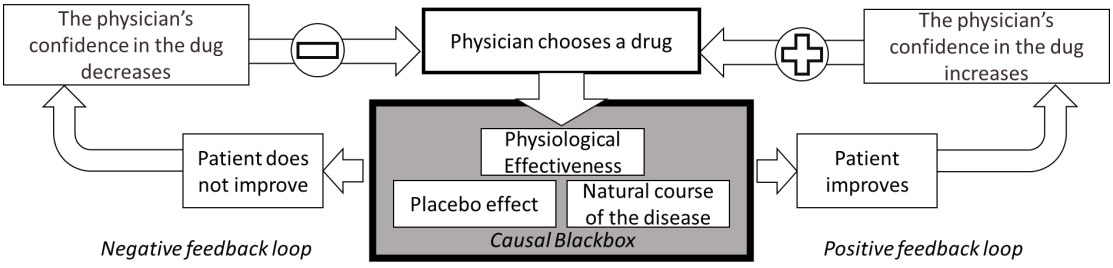
Complicating Causes

As described above, the effectiveness of the drug is not the only cause that can result in patient improvement. There are at least two other causes, namely, the placebo effect and the natural course of disease. We will sometimes refer to the placebo effect and the natural course of disease as 'complicating causes.' The significance of these complicating causes depends on the nature of the disease and varies in different scenarios. When they can play a considerable role, these complicating causes interfere with the ability to draw correct inferences about the effectiveness of a drug. Returning to the example of slot machines, suppose that every time you try one of the machines, someone secretly puts some coins in the coin hopper (the container where the payout coins are delivered). This would mislead your inference about the true reward

probability of the slot machines and misguide you to wrongly favour one of the machines that in fact might not be better than others.

This is similar for clinical experience. After the physician prescribes a drug for the patient, the patient might show improvement not because of the effectiveness of the drug, but because of placebo effect or the natural course of disease. The physician, however, might attribute the improvement to the drug and therefore overestimate the effectiveness of the drugs and wrongly favour some the dugs over others (Figure 1). Unlike the formal aspect of the physician's inference, the physician's insensitivity to the two complicating causes is non-ideal.

Figure 1: Depiction of our model of causal inference based on clinical experience.



Each time that the physician chooses a drug for a patient, the patient might subsequently show improvement because of three independent causes, represented by three random draws in the simulation: the placebo effect (L), the physiological effectiveness of drug i (P_i), and the natural course of disease (C). The probabilities associated with each of these random draws are parameters that we set in the simulations and vary from one scenario to the next.

The probability of improvement by the natural course of disease is initially set to C, and it increases by C increments per prescription up to C^{cap} over consecutive prescriptions for each patient. It then gets reset to the value of C for the next patient.

Simulations

We simulate causal inferences about effectiveness of drugs over the course of treating 100 consecutive patients. Our model physicians follow the Thompson

sampling process. At the end of the simulation, we compare each physician's subjective probability of each drug's effectiveness with their actual probability of effectiveness. This allows us to observe if physicians wrongly favour or disfavour any drugs.

We also count how many times each drug is prescribed over the course of a physician seeing 100 patients. This shows the practical manifestation of the physician's inference. Further, for each patient in each scenario, we count the number of visits to her physician (or number of prescriptions) that are required until the patient shows improvement.

We run simulations under eight scenarios, each with its own set of four parameters: physiological effectiveness of each drug i (P_i), the placebo effect (L), chance of improvement over the natural course of disease (C), and a maximum for the chance of improvement by the natural course of disease (C^{cap}). We fix the number of drugs to choose from to ten and repeat each simulation for twenty physicians. Each of P_i, L, and C are probabilities that a patient would improve due to the respective cause (physiological effectiveness, placebo, and natural course) over a defined temporal period. The temporal period is arbitrary for Scenarios 1 and 2, and for Scenarios 3-8 the temporal period is based on empirical considerations or other background knowledge of the various scenarios that allow us to inform, as realistically as possible, the values of P_i, L, and C.

The two first two scenarios are primarily for evaluating whether, in the absence of complicating causes, Thompson sampling is a good process for inferring the effectiveness of the drugs. In scenarios three to eight, we add the complicating causes and change the parameters to reflect different types of diseases and treatments. Our eight scenarios are as follows (summarized in Table 1).

Scenario 1: Ten drugs with different effectiveness, no complicating causes

In Scenario 1, the placebo effect (L) and the chance of improvement over the natural course of disease (C) are set to zero, and the ten drugs have actual effectiveness of $\{0, 0.1, 0.2, \ldots, 0.9\}$. The aim of this simulation was to determine if, in the absence of complicating causes, causal inference based on clinical experience can identify the most effective of a class of drugs.

Scenario 2: Ten drugs with equal effectiveness, no complicating causes

In Scenario 2, the placebo effect (L) and improvement over the natural course of disease (C) are set to zero, as they were in Scenario 1, but in this scenario all ten drugs have equal actual effectiveness of 0.5. The aim of this simulation was to determine if, in the absence of complicating causes, causal inference based on clinical experience can approximately discern that a class of equivalently effective drugs are indeed equivalent, and to determine if causal inferences based on clinical experience in the absence of complicating causes can approximate the actual effectiveness of the drugs.

Scenario 3: Antibiotics for bacterial infection

Scenario 3 simulates the treatment of a disease with a class of highly effective drugs, in which the disease is completely not placebo-responsive—think of the treatment of an antibiotic sensitive bacterial infection with a strong antibiotic. The defined temporal period is, say, one week: it is a bad bacterial infection and the probability of improving due to the natural course of disease is 5% during that week, though proper administration of antibiotics would cure the infection within the week 90% of the time. Thus, the placebo effect (L) is set to zero, and improvement due to the natural course of disease (C) is set to 0.05 and the maximum improvement due to the natural course of disease (C^{cap}) is also set to 0.05, and all the drugs have an equal effectiveness of 0.9.

Scenario 4: Antidepressants, champion's view

Scenarios 4 and 5 simulate the treatment of depression with antidepressants from two perspectives on the effectiveness of antidepressants. Whether antidepressants have clinically relevant physiological effects is debated (Moncrieff and Kirsch 2015; Munkholm, Paludan-Müller, and Boesen 2019). The basis for this debate is how to interpret the small but non-zero effect sizes observed in meta-analyses of antidepressants (Cipriani et al. 2018). The perspective we call the "champion's view" holds that we should interpret the small but non-zero effect sizes on their face, thereby maintaining that antidepressants indeed have a small but positive average effectiveness. The perspective we call the "sceptic's view" holds that those small non-zero effect sizes can be best explained by biases in the relevant trials; one bias in particular that some sceptics note is 'blind-breaking', which is the empircally substantiated phenomenon whereby subjects in the drug group of trials on antidepressants accurately guess which group they are in based on their experience of side effects, thereby exaggerating the placebo effect in that group (Gøtzsche 2014).

Scenario 4 simulates the case from the champion's perspective. The temporal period is around six weeks, roughly the duration of a typical trial of antidepressants. In meta-analyses of these trials a notion of 'responder' is often used; a responder is a subject whose depression severity score goes down by at least 50% compared with their pre-trial depression severity. The parameters for the chance of being a responder due to the drugs or placebo are based on data reported in (Cuijpers, Stringaris, and Wolpert 2020), in which roughly 60% of subjects in the drug group are responders and roughly 40% of subjects in the placebo group are responders (see Tabatabaei Ghomi and Stegenga, 2022 for further discussion and critique of the use of such responder analyses). Moreover, some empirical evidence suggests that if left untreated, on average about 50% of depressed patients will improve over the course of one year, or roughly 5% of patients will improve every six weeks (see Cuijpers, Stringaris, and Wolpert 2020; Posternak and Miller 2001). Thus, the placebo effect (L) is set to 0.4, improvement due to the natural course of disease (C) is set to 0.05, maximum improvement due to the natural course of disease (C^{cap}) is to 0.5, and all the drugs have an equal effectiveness of 0.2.

Scenario 5: Antidepressants, sceptic's view

Scenario 5 simulates antidepressants from the sceptic's perspective, who holds that the improvements observed in trials of antidepressants are not due to the physiological effectiveness of the drug but are merely due to placebo effects and methodological biases. In this simulation the chance of observing improvement due to the effects of the drugs is set to zero, and the rest of the parameters are the same as in Scenario 4. This scenario can also represent the use of many treatments in complementary and alternative medicine, in which the treatments are probably ineffective but there is a high chance that the peculiar practices of complementary and alternative medicine result in significant placebo effects. All of the parameters in Scenario 5 are that of Scenario 4, except for the drug effectiveness, which is set to 0.

Scenario 6: Treatment of the common cold

Scenario 6 simulates a disease such as the common cold, for which we do not have a highly effective drug, yet the patient rapidly gets well by the natural course of disease. The temporal period in this scenario is one day. Here we are supposing that common colds clear up in around five days, and that a patient tries a new treatment every day. The probability that a patient would improve in any given day increases by 0.2 every

day ($C = 0.2$, $C^{cap} = 1$). We also suppose that any intervention would be nearly useless, though there may be some interventions with very modest effectiveness, and placebo response is also minimal. Thus, the physiological effectiveness (P_i) and placebo effect (L) are both set to 0.05.

Scenario 7: Antibiotic-resistant infection

Scenario 7 simulates a disease such as treatment of an antibiotic-resistant infection, which are particularly difficult to treat, and for which there is little chance that the patient gets better by placebo effects or the natural course of disease. Accordingly, the physiological effectiveness (P_i) and placebo effect (L), as well as the chance of improvement by the natural course of diseases (C) and the maximum natural course of improvement (C^{cap}) are all set to 0.05.

Scenario 8: Context-sensitive treatments

Scenario 8 has a slightly different approach compared to other scenarios. Here we set L and C to zero, assuming no placebo effect or improvement by natural course of disease. The value for the physiological effectiveness of each drug i (P_i) is randomly picked at the time of each prescription from a uniform distribution over 0.2 and 0.8. So, in each prescription, the same drug might be very effective (P_i close to 0.8), only mildly effective (P_i close to 0.2), or somewhere in between. Nonetheless, the average value of P_i for all drugs is 0.5. The scenario simulates context-sensitive drugs. Examples are drugs that are sensitive to the genetic background of patients, have high interactions with food or other drugs, or some cases of alternative medicine where the effect is claimed to depend on many factors besides the drug itself.

Table 1: Summary of parameter values for our eight clinical scenarios.

Parameter	Scenario 1 Variable effectiveness, no complicating causes	Scenario 2 Same effectiveness, no complicating causes	Scenario 3 Antibiotics for bacterial infection	Scenario 4 Antidepressants: champion's view	Scenario 5 Antidepressants: skeptic's view	Scenario 6 Treatment of the common cold	Scenario 7 Antibiotic-resistant infection	Scenario 8 Context-sensitive drugs
effectiveness of drug i (P_i)	{0, 0.1, 0.2, …, 0.9}	0.5	0.9	0.2	0.0	0.05	0.05	[0.2, 0.8]
placebo effect (L)	0	0	0	0.4	0.4	0.05	0.05	0
course of disease (C)	0	0	0.05	0.05	0.05	0.2	0.05	0
course max (C^{cap})	0	0	0.05	0.5	0.5	1	0.05	0

3. Results and discussion

Figure 2 summarises the results of the simulations for each of the simulated scenarios, for a single randomly chosen physician. Each of the sub-figures shows the probability density function representing the physician's confidence that each drug is effective after her 'career' of treating 100 patients. Vertical dotted lines show the actual probability of effectiveness of the drugs (P_i). Figure 3 displays the proportion of times this randomly chosen physician chose the ten drugs, over the duration of their career. The patterns depicted in Figures 2 and 3 are from one random physician in each scenario, but the general observations were robust over the careers of all twenty physicians under all scenarios and those results are provided in supplementary material available online. Thus, Figure 2 depicts the inference stage of our model while Figure 3 depicts the drug choice stage of our model. Also, we have created animations of the proportion of times all twenty physicians have chosen each drug, at every point in the career of each physician, for all eight scenarios (these can be provided on request and made available as supplementary online material). Figure 4 visually displays the minimum, maximum, and average number of prescriptions that are required for a patient to show improvement, for each physician in each scenario, while Table 2 shows the average number of prescriptions per patient in each scenario.

Figure 2: A physician's probability density function for the ten drugs, representing their estimates of effectiveness, after 100 patients, under different scenarios. Vertical dotted lines indicate P_i

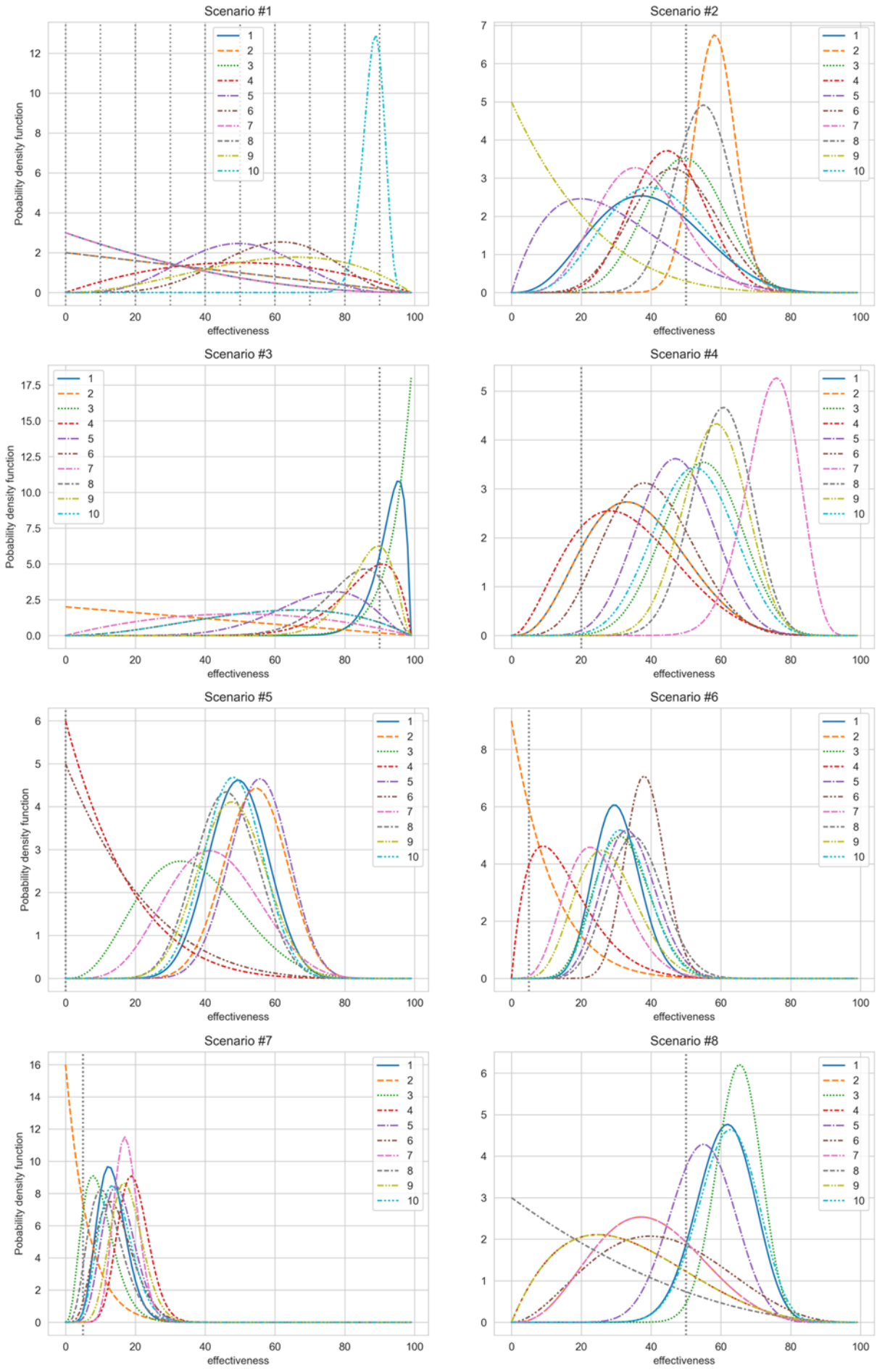


Figure 3: The number of times a physician has prescribed any of the ten drugs (1 to 10) over 100 patients, under different scenarios.

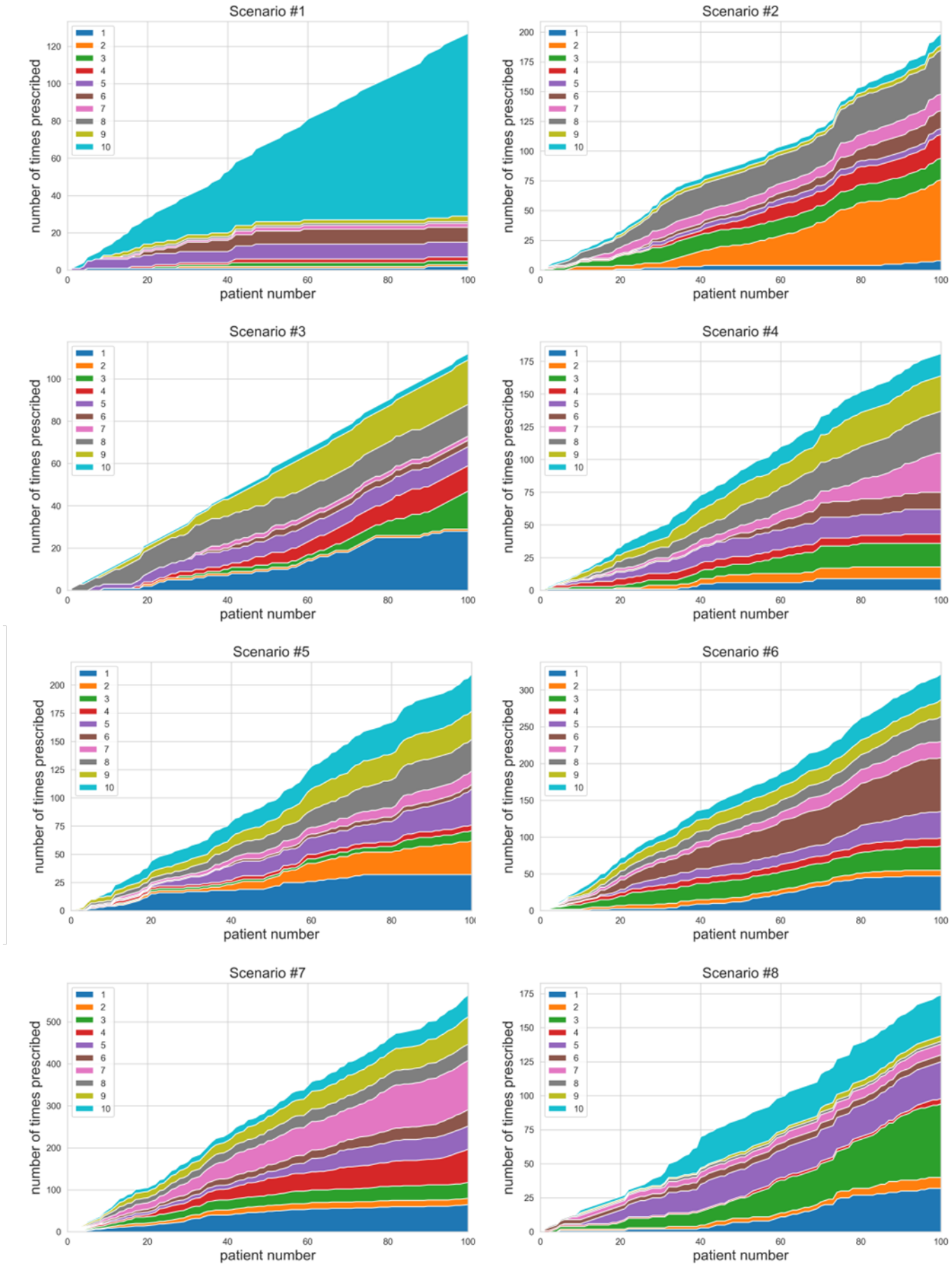


Figure 4: The minimum, maximum, and average number of prescriptions before a patient shows improvement, under different scenarios.

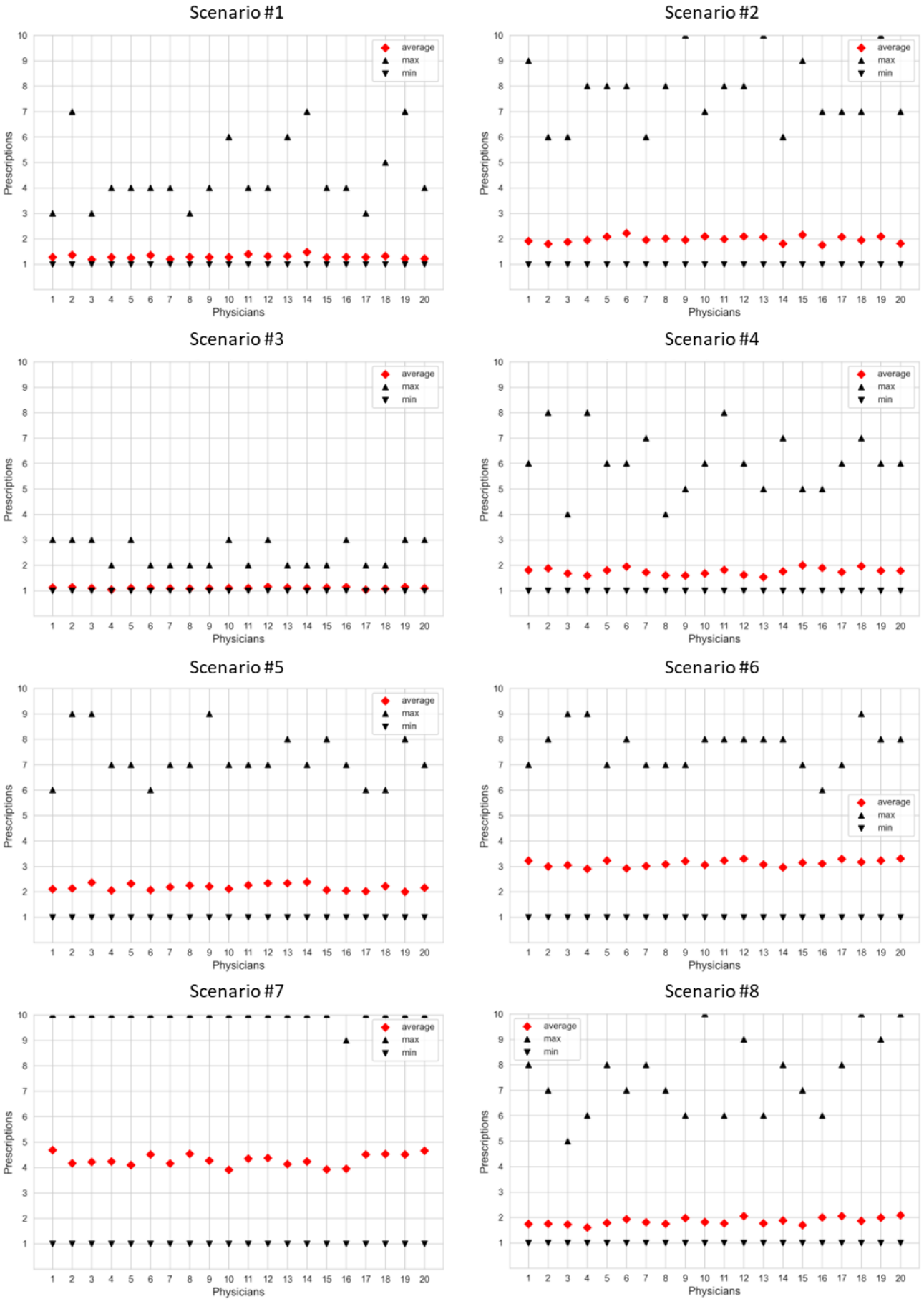


Table 2. Average number of prescriptions per patient

Scenario	Prescriptions per patient
1	1.3
2	2.0
3	1.1
4	1.8
5	2.2
6	3.1
7	5.6
8	1.9

Scenario 1: Ten drugs with different physiological effectiveness, no complicating causes

The results of Scenario 1 shows that in the absence of a placebo effect or natural course of disease, causal inference based on clinical experience can identify the most effective drug among ten drugs of differing effectiveness. As shown in Figure 2, the physician correctly identifies the most effective drug, and is very certain about its effectiveness (narrow distribution). She has less accurate inferences about the effectiveness of the other drugs, but because she has reliably found the most effective drug, ascertaining the effectiveness of the other drugs is of little clinical value for the physician or her patients. As shown in Figure 3, the physician quickly identifies the most effective drug and prescribes that drug for most patients. All patients improve in this scenario and patients need only an average of 1.3 prescriptions to improve (Figure 4 & Table 2).

Scenario 2: Ten drugs with equal physiological effectiveness, no complicating causes

Scenario 2 shows that in the absence of complicating causes and working with equivalent drugs, a physician can make only modestly reliable causal inferences based on clinical experience. As Figure 2 shows, in this scenario the physician does not come to a confident and very accurate inferences about the effectiveness of all the drugs, and strongly favours some drugs over others (Figure 3). Nonetheless, she has an approximately correct inference about the effectiveness of many of the drugs. Nearly all patients improve in this scenario and patients need only an average of ~2.0 prescriptions to improve (Figure 4 & Table 2).

In summary, the results of scenarios 1 and 2 show that in the absence of complicating causes, clinical experience can be a good guide to the actual effectiveness of the drugs, particularly if the drugs considerably differ in their actual effectiveness. Yet Figure 3 shows that the differences in physicians' estimates of the effectiveness of the various drugs can be large enough such that in practice the physicians choose some of the drugs much more often compared to others (while Figure 3 shows the results for a single physician, this was a pattern observed across the group of physicians).

Scenario 3: Antibiotics for bacterial infection

As Figure 2 shows, causal inference based on clinical experience is quite reliable in this scenario. When the drugs are very effective, and the complicating causes are not strong, such inferences can be reliable. However, the physician has much greater confidence in their estimates of the effectiveness of some drugs, despite the fact that all the drugs in this scenario are equally effective, and in particular, physicians underestimated the effectiveness of some of the drugs. Physicians subsequently tended to prescribe some drugs in this scenario much more often than others, as shown in Figure 3. So, while inferences about effectiveness were relatively reliable in this scenario, there was nevertheless unwarranted favouring of some drugs over others. All patients improve in this scenario and patients need only an average of 1.1 prescriptions to improve (Figure 4 & Table 2), which is of course intuitive given how uniformly effective the drugs are.

Scenario 4: Antidepressants, champion's view

Figure 2 shows that the physician's inferences about the effectiveness of the drugs can be as much as ~60% higher than the actual effectiveness of the drugs, when the drugs are moderately effective and the placebo effect is strong. The inference about the effectiveness of each particular drug has a relatively narrow distribution, meaning that the physician is confident about this inaccurate inference. Figure 2 also shows that the physician wrongly infers very different effectiveness of drugs that have completely equivalent effectiveness (for example, compare drug 4 to drug 7). Physicians also tended to prescribe some drugs in this scenario more often than others in this scenario, as seen in Figure 3. In short, this scenario displays two fallacies of inference being committed to a very large degree: overestimation of effectiveness and unwarranted favouring. All patients improve in this scenario and patients need only an average of 1.8 prescriptions to improve (Figure 4 & Table 2), which could be seen as surprising given the relatively low effectiveness, though this is explained by the strong placebo effects.

Physicians were able to roughly estimate the overall probability of improvement due to all causes, in this scenario and in Scenario 5, though they were unable to distinguish improvement due to physiological effects from placebo effects. In ongoing development of the model we are adding some capacity to discriminate between expected effects of the three causes.

Scenario 5: Antidepressants, sceptic's view

The physician's inferences in this scenario—that is, in which the drugs are ineffective but placebo effect is strong—are very inaccurate (Figure 2). The physician deems most of these completely ineffective drugs to be quite effective (up to ~60%). Also, just as in Scenario 4, the physician thinks that there are relatively large differences in the effectiveness of these drugs, despite the fact that they are all equally ineffective (for example, compare drugs 3 and 5). In this scenario physicians also had favourite drugs, tending to prescribe some drugs in this scenario much more often than others (Figure 3). All patients improve in this scenario and patients need only an average of 2.2 prescriptions to improve (Figure 4 & Table 2); this result is impressive given that the drugs in this scenario are completely ineffective, and again, this result is best explained by the strong placebo-responsiveness of the disease being treated in this scenario.

Scenario 6: Treatment of the common cold

The physician's inferences in this scenario—namely, situations involving drugs of low effectiveness but a strong natural course of disease—are also very inaccurate (Figure 2). The physician wrongly deems most of these ineffective drugs to be effective (up to ~40%). Also, as in scenario 5, the physician thinks that there are relatively large differences in the effectiveness of these drugs, despite the fact that they are all equally (and minimally) effective (for example, compare drugs 4 and 6). Like in other scenarios, this scenario saw physicians prescribing some drugs much more frequently than others, thereby displaying unwarranted favouring of some drugs (Figure 3). All patients improve in this scenario and patients need an average of 3.1 prescriptions to improve (Figure 4 & Table 2); this should not be very surprising, since virtually everyone can shake the common cold.

Scenario 7: Antibiotic-resistant infection

Figure 2 shows that in this scenario, involving the treatment of relatively intractable diseases, the physician can reliably infer that the drugs have very low effectiveness. Nonetheless, the physician slightly overestimates the actual effectiveness of the drugs due to the complicating causes. Similar to the other scenarios, the differences in the inferred effectiveness of the drugs are large enough to make the physician favour some drugs over their equivalents. Many patients do not improve in this scenario, and on average patients require 4 to 5 prescriptions to improve (Figure 4).

Scenario 8: Context-sensitive drugs

Figure 2 shows that context-sensitivity of drugs can confuse the physician's inference even in the absence of complicating causes. Similar to the results of Scenario 2, the physician infers different values for effectiveness of equivalent drugs and ends up strongly favouring some drugs over others (Figure 3), despite the fact that the average effectiveness of all the drugs is the same. Most patients get better when given enough prescriptions in this scenario (Figure 4), and on average patients needed ~2 prescriptions to get better (Table 2).

4. Conclusion

In our model, when causes such as the placebo effect and improvement due to the natural course of disease play a significant role in patient improvement, or when drugs are context-sensitive, causal inference based on clinical experience is unreliable. These inferences are more reliable when the disease being treated is not placebo-responsive and has little or no natural course of improvement, especially when some drugs are significantly better than others, or when all of the drugs are highly effective. From a patient's perspective making such inferences are equally challenging, since virtually all patients improve when given enough time or enough prescriptions, even when the drugs are completely ineffective.

Thus, the evidence-based medicine position, which doubts the reliability of physicians' causal inferences based on clinical experience, is partially vindicated by our results. At least, for diseases with some natural course of improvement or some placebo-responsiveness, such inferences are unreliable in our model. Yet, for other diseases which have little placebo-responsiveness and little natural course of improvement, such inferences can be relatively reliable.

However, both sides of the debate regarding the reliability of such inferences will have reasons to assess our results cautiously. A defender of the evidence-based medicine position could note that one significant idealisation in our approach is that physicians in our model update their credences and choose among the available drugs as an ideal reasoner would, which is unrealistic, since we have good reasons to think that physicians, like the rest of us, are liable to various sorts of reasoning biases. It is those reasoning biases which motivate the methodological strictures of the evidence-based medicine movement in the first place: testing the effects of drugs with randomised trials rather than with reference to clinical experience is intended to block the impact of those reasoning biases. On the other hand, a defender of the use of clinical experience to test the effects of drugs could note that another significant idealisation in our approach is that physicians in our model are not able to modulate their inferences based on the extent to which they believe the disease being treated is placebo-responsive or has a natural course of improvement, and this too is unrealistic, since it is plausible to think that physicians do in fact modulate such inferences in that manner. Both such considerations are apt and warrant caution in interpreting our results. Yet one general remark in response to these considerations is to note that they pull in different directions. The champion of evidence-based medicine would be in effect claiming that our model does not sufficiently represent just how unreliable clinical inferences can be, while the champion of clinical experience would be in effect claiming that our model does not sufficiently represent just how reliable clinical inferences can be. The two kinds of idealisations that these two considerations appeal to would have opposing implications for how to interpret our results, and which of the considerations is weightier remains an open question.

In future work we aim to address that question. We plan to model various reasoning biases, such as confirmation bias, novelty bias, and other reasoning patterns such as risk aversion. As noted, in the present model the physicians do not modulate their inferences based on prior beliefs about how placebo-responsive or naturally-improving a disease is, and in future work we plan to add such complexity to our model. Moreover, the extent to which real physicians modulate their inferences based on prior beliefs about the placebo response and the natural course of improvement is itself an empirical question, and in current research we are collaborating with colleagues performing experimental work to illuminate that.

Our model is of course focused solely on causal inference from clinical experience. Yet, clinical expertise is not merely about inferring the effects of drugs, but rather is about a much wider range of phenomena, including diagnosis, understanding relevant

details of particular patients, and the importance of practical know-how. Expertise is, of course, a complex and multidimension property, and our results are focused entirely on one angle, namely the reliability of causal inferences made by physicians based on their routine clinical experience.

Moreover, the results presented here are a small sample of the range of possible results to be gleaned from our model. We simulated two theoretical scenarios as model checks and six scenarios that we take to be realistic depictions of routine clinical practice. Yet, our model permits the tuning of many combinations of parameters, and in future work we plan to more fully explore the parameter space, modelling other kinds of clinical scenarios. In the present model patient improvement is dichotomous, and we can instead operationalise patient outcomes as a continuous property (a graded measure of symptom severity, say).

One might think that a champion of evidence-based medicine could respond to our results by noting that our method involves a physician seeing a population of patients, and so the physician's evidence could approximate the population-level evidence that a randomised trial provides, in which case any scenario that suggests some reliability to causal inference from clinical experience (i.e. Scenario 7) can be explained by the population-level structure of the physician's evidence. Yet, given that the set of evidence that a particular physician acquires in our model comes from a method which is non-randomised, unblinded, and uncontrolled, the evidence is very far from the gold-standard that evidence-based medicine stipulates. Indeed, the evidence hierarchies of evidence-based medicine place expert opinion and clinical judgement at the bottom, and that is regardless of whether such judgement is a function of experience gleaned from a long sequence of patient interactions. The scenarios that suggest some reliability to causal inference from clinical experience are best explained by the absence of strong complicating causes such as placebo or natural course of improvement, rather than the fact that the experience of our simulated physicians is constituted by a population of patients.

Despite the simplicity of the present model, the preliminary results are striking. We show that causal inference from clinical experience can be reliable under particular conditions, but in other conditions, exemplifying very common real clinical scenarios, physicians wildly overestimate drug effectiveness and wrongly favour some drugs over equivalent drugs. Given the noted limitations of our model, these results must be interpreted with caution, yet they are an initial step toward a deeper understanding of this ubiquitous mode of causal inference.

References

Blunt, Christopher. 2015. "Hierarchies of Evidence in Evidence-Based Medicine." London School of Economics and Political Science.

Bouneffouf, Djallel, Irina Rish, and Charu Aggarwal. 2020. "Survey on Applications of Multi-Armed and Contextual Bandits." In *2020 IEEE Congress on Evolutionary Computation (CEC)*, 1–8.

Cartwright, Nancy. 2017. "How to Learn about Causes in the Single Case." *Durham University: CHESS Working Paper No* 4: 2017.

Choudhry, Niteesh K, Robert H Fletcher, and Stephen B Soumerai. 2005. "Systematic Review: The Relationship between Clinical Experience and Quality of Health Care." *Annals of Internal Medicine* 142 (4): 260–73.

Cipriani, Andrea, Toshi A Furukawa, Georgia Salanti, Anna Chaimani, Lauren Z Atkinson, Yusuke Ogawa, Stefan Leucht, et al. 2018. "Comparative Efficacy and Acceptability of 21 Antidepressant Drugs for the Acute Treatment of Adults with Major Depressive Disorder: A Systematic Review and Network Meta-Analysis." *Lancet* 391: 1357–66.

Cuijpers, Pim, Argyris Stringaris, and Miranda Wolpert. 2020. "Treatment Outcomes for Depression: Challenges and Opportunities." *The Lancet Psychiatry* 7 (11): 925–27.

Dewitt, Barry, Johannes Persson, Lena Wahlberg, and Annika Wallin. 2021. "The Epistemic Roles of Clinical Expertise: An Empirical Study of How Swedish Healthcare Professionals Understand Proven Experience." *Plos One* 16 (6): e0252160.

Gøtzsche, Peter C. 2014. "Why I Think Antidepressants Cause More Harm than Good." *The Lancet Psychiatry* 1 (2): 104–6.

Grünbaum, Adolf. 1981. "The Placebo Concept." *Behaviour Research and Therapy* 19 (2): 157–67.

———. 1986. "The Placebo Concept in Medicine and Psychiatry." *Psychological Medicine* 16 (1): 19–38.

Guyatt, Gordon, John Cairns, David Churchill, Deborah Cook, Brian Haynes, Jack Hirsh, Jan Irvine, et al. 1992. "Evidence-Based Medicine: A New Approach to Teaching the Practice of Medicine." *JAMA* 268 (17): 2420–25.

Healy, David. 2011. "Science, Rhetoric and the Causality of Adverse Events." *International Journal of Risk \& Safety in Medicine* 23 (3): 149–62.

———. 2020. "Medical Nihilism by Jacob Stegenga: Is Operationalism the Answer to Nihilism?" *Studies in History and Philosophy of Biological and Biomedical Sciences* 81: 101272.

Heesen, Remco. 2018. "When Journal Editors Play Favorites." *Philosophical Studies* 175 (4): 831–58.

Holman, Bennett, and Justin P Bruner. 2015. "The Problem of Intransigently Biased Agents." *Philosophy of Science* 82 (5): 956–68.

Howick, Jeremy. 2011. *The Philosophy of Evidence-Based Medicine*. John Wiley \& Sons.

———. 2017. "The Relativity of 'Placebos': Defending a Modified Version of Grünbaum's Definition." *Synthese* 194 (4): 1363–96.

Kuleshov, Volodymyr, and Doina Precup. 2014. "Algorithms for Multi-Armed Bandit Problems." *ArXiv Preprint ArXiv:1402.6028*.

Lai, Tze Leung. 1987. "Adaptive Treatment Allocation and the Multi-Armed Bandit Problem." *The Annals of Statistics*, 1091–1114.

Mayo-Wilson, Conor, and Kevin J S Zollman. 2021. "The Computational Philosophy: Simulation as a Core Philosophical Method." *Synthese*, 1–27.

Meehl, Paul E. 1986. "Causes and Effects of My Disturbing Little Book." *Journal of Personality Assessment* 50 (3): 370–75.

Moncrieff, Joanna, and Irving Kirsch. 2015. "Empirically Derived Criteria Cast Doubt on the Clinical Significance of Antidepressant-Placebo Differences." *Contemporary Clinical Trials* 43: 60–62.

Munkholm, Klaus, Asger Sand Paludan-Müller, and Kim Boesen. 2019. "Considering the Methodological Limitations in the Evidence Base of Antidepressants for Depression: A Reanalysis of a Network Meta-Analysis." *BMJ Open* 9 (6): e024886.

Posternak, Michael A, and Ivan Miller. 2001. "Untreated Short-Term Course of Major Depression: A Meta-Analysis of Outcomes from Studies Using Wait-List Control Groups." *Journal of Affective Disorders* 66 (2–3): 139–46.

Rubin, Hannah. 2022. "Structural Causes of Citation Gaps." *Philosophical Studies* 179 (7): 2323–45.

Šešelja, Dunja. 2022. "Agent-Based Models of Scientific Interaction." *Philosophy Compass* 17 (7): e12855.

Stegenga, Jacob. 2018. *Medical Nihilism*. Oxford University Press.

Tabatabaei Ghomi, Hamed, and Jacob Stegenga. 2022. "Conventional Choices in Outcome Measures Influence Meta-Analytic Results." *Philosophy of Science* 89 (5): 949–59.

Tonelli, Mark R, and Devora Shapiro. 2020. "Experiential Knowledge in Clinical Medicine: Use and Justification." *Theoretical Medicine and Bioethics* 41 (2): 67–82.

Villar, Sofía S, Jack Bowden, and James Wason. 2015. "Multi-Armed Bandit Models for the Optimal Design of Clinical Trials: Benefits and Challenges." *Statistical Science: A Review Journal of the Institute of Mathematical Statistics* 30 (2): 199.

Weatherall, James Owen, and Cailin O'Connor. 2021. "Conformity in Scientific Networks." *Synthese* 198 (8): 7257–78.

Zollman, Kevin J S. 2007. "The Communication Structure of Epistemic

Communities." *Philosophy of Science* 74 (5): 574–87.

———. 2015. "Modeling the Social Consequences of Testimonial Norms." *Philosophical Studies* 172: 2371–83.