DISCUSSION NOTE

The Reference Class Problem and Probabilities in the Individual Case: A Response to Fuller

Arjun Devanesan

Department of Philosophy, King's College London, London, UK
Email: adevanesan@doctors.org.uk

Abstract

In a recent article on the interpretation of probability in evidence-based medical practice, Jonathan Fuller argues that we should interpret probabilities as credences in individual cases because this avoids some important problems. In this article, I argue that Fuller misidentifies the real issue and so fails to offer a meaningful solution to it. The real problem with making probability judgments in individual cases is deciding which objective considerations ought to constrain our formation of credences. This leads us to the reference class problem, which, as Alan Hajek argues, is a problem for any interpretation of probability.

1. Introduction

What could it mean for a doctor to think (or say) that his patient has a 10% chance of having a stroke in the next 5 years? Even further, what could it mean for him to think that starting a statin will reduce that chance? This is the crux of the question taken up by Jonathan Fuller in his recent article on the interpretation of probability (or risk) in the individual case (Fuller 2020). In this article, Fuller focuses on how one ought to interpret probabilities involving individuals, and he does this by comparing the ontic propensity theory and the epistemic credence theory. He concludes that credences are the most sensible way of interpreting individual probabilities and argues that this allows evidence of risk to seamlessly incorporate clinical experience and intuition.

Although Fuller is correct that there is a problem with making probability judgments about individual cases, the problem is not solved by simply interpreting probabilities as credences instead of propensities. As we shall see, the real problem is not a choice between different interpretations of probability at all but rather understanding which objective considerations ought to constrain our formation of credences.

In this article, I distinguish probability, propensities, and credences, following the approach suggested by Paul Humphreys (1989, 2004, 2019) and Mauricio Suárez (2018). I then argue that the real problem for inferring probabilities in the individual case is the reference class problem. Hajek (2007) has shown that this is a problem for any

interpretation of probability, and so Fuller's endorsement of a subjective theory of probability in individual cases does not tackle the real difficulty in inferring probabilities in individual cases.

2. Fuller's argument, from propensities to credences

In Fuller's article, he attempts to "show how EBM's [evidence-based medicine's] view of evidence and risk makes the problem of the meaning of population evidence difficult to resolve, and [he] propose[s] an epistemic interpretation of individual patient probabilities as a corrective" (Fuller 2020, 1121). Fuller examines the two dominant philosophical interpretations of probability, the ontic propensity theory and the epistemic credence theory. He relates a number of well-known problems with inferring single-case propensities and argues that these are avoided by interpreting these probabilities as credences.

2.1. Ontic probability

Ontic probabilities are facts about processes that produce outcomes or events in the world. Propensities are the dispositional properties of these processes (also known as *chance setups*) to produce particular outcomes and were thought by Popper et al. to properly represent single-case ontic probabilities (Gillies 2000). Fuller uses the terms *risk*, *chance*, and *probability* interchangeably, but I will use the term *probability* throughout to refer only to a formal concept in certain mathematical theories (e.g., von Mises's theory of collectives), which we ought to distinguish from worldly properties called *propensities.* Furthermore, although probabilities are sometimes interpreted as propensities (and vice versa), there are good reasons not to equate the two (Humphreys 2004; Suárez 2018). Also, I will use the term *credence* rather than *subjective probability* to distinguish the former as a doxastic attitude toward events in the real world as opposed to properties of the world.

Humphreys (2019) in particular distinguishes two general approaches to the "interpretation of probability." On the one hand, one might provide an explicit definition of probability, such as Popper's propensity theory, which identifies probabilities with dispositions of repeatable conditions to produce outcomes in a certain frequency (see Gillies 2000, 114–18). On the other hand, Humphreys's (2019) suggestion is to take probability as a primitive term in a formal theory, implicitly defined by the axioms that make up that theory. The virtue of this approach is that it clearly distinguishes mathematical theories of probability (such as Kolmogorov's measure-theoretic system) from probabilistic models. In such models, for example, particular probability distributions are substituted for P in the probability space $<\Omega, I, P>$ in order to represent some particular physical system in the world. These models can then be clearly distinguished from the real systems that are being represented.

Consider the propensity of a population to manifest a certain state of affairs, measured by the frequency that it occurs over a certain time frame. In Fuller's example, if 3% of a certain population had a heart attack when observed over a 10-year period, one would infer a propensity for having a heart attack of 3% for that population. Populations are also sometimes called *reference classes* when they are defined according to particular criteria. So, the population of men with diabetes is a reference class and may have a probability P of having a heart attack over a 10-year

period. If we want to know what *P* is, we can simply observe a population of men with diabetes and see how many of them have a heart attack in this time frame, then apply the usual statistical inferential techniques. We might then infer the propensity of a (generic) man with diabetes having a heart attack. However, what about a particular man with diabetes?

If one were to adopt von Mises's theory, the probability in the single case is undefined and so cannot be used to infer the propensity. So, propensity theory was first proposed (by Karl Popper) as a theory about single-case probability. But there are a number of serious difficulties with early propensity theories, notably the failure to show that probabilities in the individual case can be used to infer propensities because of Humphreys's paradox[1] (Gillies 2000; Humphreys 2004). Subsequent versions distinguish between long-run propensity theory and single-case propensities (Gillies 2000) in order to avoid these problems (Gillies 2000; Lyon 2014; Humphreys 2019).

Fuller says that "there is good reason to suspect that conceptually the patient's risk is different from the AR [absolute risk], the … frequency in the population" (Fuller 2020, 1122). He also points out that "we do know (e.g.) that the AR of heart attack or stroke will vary widely from less than 1% to over 50% in common subgroups of patients (Goff et al. 2014), which suggests a large spread of cardiovascular propensities in the population. We have tools to further stratify patient risk, but not down to the level of the individual. In general, we should worry that any AR and ARR [absolute risk reduction] might be a poor estimate for the individual unless we have evidence to suggest otherwise" (Fuller 2020, 1124).

He is quite right that the probability of an outcome for any individual may well be different from the average probability in a population (or reference class) to which that individual belongs. As such, if the probability of a heart attack in the reference class is 3%, that does not entail that the probability for any individual member is 3%. It may be that the probabilities for the individuals in the population are all 0 or 1 (in the case that determinism is true) but that only 3% of the individuals will experience the event. So, Fuller argues that one cannot claim that the propensity for an individual member of a reference class can be inferred from population frequencies and the standard probability calculus.

Although Fuller is quite right that different members of a population will differ with respect to whether and to what extent they are affected by the causal determinants of a particular outcome, he is not clear on exactly how this relates to probability, propensity, and credences in the individual case. Although long-run propensity theories may avoid Humphreys's paradox (e.g., Gillies 2000), they are no help in the individual case (which Gillies takes to be only interpretable as credences).

That said, it is well known that mathematical representations of a system do not necessarily correspond to the represented physical systems out in the world (Humphreys 1989). So, when formal theories of probability fail to represent the propensities of certain chance setups (e.g., the single case), they may not properly

---

[1] Humphreys's paradox arises because when we define a conditional probability (e.g., P(A|B)), such as the probability of rain on Tuesday given the weather on Monday, the inverse is also defined (i.e., P(B|A)). Whereas the first is straightforwardly interpreted as a propensity, how are we to interpret the probability of a certain weather pattern on Monday, given rain on Tuesday, as a propensity?

inform us of their behavior. However, as Humphreys argues, "rather than this being construed as a problem for propensities, it is to be taken as a reason for rejecting the current theory of probability as the correct theory of chance" (Humphreys 2019, 143).

Following Humphreys'[2] suggestion, Mauricio Suárez (2018) defends an approach for single-case propensities in which he argues that these should not be interpreted as probabilities but as probability functions that are indexed by certain generating conditions. Importantly for this theory, the generating conditions serve to define a system-specific statistical model (or system-specific probability space) rather than serving as factors for conditioning in a conditional probability function. This specificity allows for a tailored approach in individual cases. It (roughly) views probability in an individual patient as a single-case propensity, given case-specific generating conditions, which can be tested by means of an AR in an appropriate population.[3] The question is whether, or under what conditions, the frequency in the population (the AR) is a good measure of individual propensity for a member of the population.

If the population is made up of a number of identical individuals, say, a hundred tosses of the same die, then the AR would be a good measure of the propensity of the die to land on a particular face on any particular toss. The problem in medicine is that populations are heterogeneous. The AR is then only a good measure of an individual's risk to the extent that the members of the population are similar with respect to the risk profile of the individual in question. However, individuals of interest are often relevantly dissimilar to the populations for which we have epidemiological data, and this is the real source of the difficulty in estimating their propensity for a certain outcome. As we shall see, interpreting the AR as a credence does nothing toward solving this problem.

2.2. Credences

Credences are a degree of belief in a proposition, and in order to measure credences, they are often taken to be equivalent to betting preferences. A degree of belief stated by the proposition $C(E)$ is measured by the lowest betting quotient q a person would propose if, given a stake S in the event E, qS is the payoff if E occurs. From this simple principle, both Ramsey and De Finetti showed that as long as your bets obey the axioms of probability, your credences are coherent, and you are never guaranteed to lose a series of bets based on them (Gillies 2000).

For a radical subjectivist like De Finetti, coherence is the only constraint on formulating credences. You may have any credences you want as long as they are coherent according to the axioms of probability theory. As such, it would be coherent to believe that the chance of the sun exploding tomorrow is 0.99 as long as you believe that the chance of it not exploding is 0.01, irrespective of the evidence. This pair of credences may be perfectly coherent, but I struggle to see how they could be considered rational. Moreover, such radically permissive credences are immune to

---

[2] See also Gillies (2000, chap. 6 and 7) for the suggestion that we add to the standard probability space $<\Omega, I, P>$ a set of repeatable conditions, S.

[3] My thanks to an anonymous reviewer for this point.

evidence and, as such, need not reflect how the world is. So, they can hardly be a guide to life, let alone medical decision making.

For credences to be rational and a guide to life, they must be further constrained in some way. First and foremost, as David Lewis (1980) argued, a rational agent would calibrate her credence to the single-case probability if she knew it. Lewis called this the *principal principle*, and it is one of a number of chance credence norms stating that credences should be calibrated to known probabilities. The principal principle states that a credence should match the probability of A holding given by proposition X and any admissible evidence E: $C(A|XE) = x$ (Lewis 1980). So, when considering an individual case, a rational agent would calibrate her credence to the single-case probability if she knew it. However, this is not practical. As Lewis argues, for a single-case probability, "the chance distribution at a time and a world comes from any reasonable initial credence function by conditionalizing on the complete history of the world up to the time, together with the complete theory of chance for the world" (Lewis 1980, 277). How often are we able to conditionalize on the complete history of the world?

A similar but more practical principle is Beebee and Papineau's (1997) relative principle: the correct credence for an agent to attach to outcome $\{o_i\}$ is the probability of $\{o_i\}$ relative to the agent's knowledge of the setup. This credence may only ever approximate the true probability, but that does not make the credence irrational. If all that is known is the frequency in a reference class, then Fuller is right that it would be rational to calibrate your credence to this for any individual member of the class. As such, Fuller argues that although the AR of an event in a reference class may not tell us the propensity for an individual member of the class, a rational agent would set her credence equal to the AR for the reference class if that was all she knew.

3. The reference class problem

That said, a rational credence must be calibrated to *everything* that is known about the individual. As Fuller rightly argued, we often have good reasons to suspect that the AR in one particular reference class is a poor estimate of the propensity for one of its members. But in any particular case, we must, at a minimum, take a view on whether we think that the propensity in the individual case is equal to the AR in the reference class of which it is a member based on our total evidence. If the AR in a reference class is all we know, then we cannot do any better than inferring the propensity of an individual member of the class to be equal to the AR. This is why it is sometimes rational to set your credence in an individual propensity equal to the AR.

However, I may well suspect that the probability of an event for an individual is different from the average in a reference class of which it is a member because I know that the individual is a member of other reference classes in which the probability is different. A rational agent then faces the problem of calibrating her credence to not one but many probabilities from multiple reference classes, and the frequency of an outcome given the reference class of everything that is known about the individual is not available.

This gives rise to the reference class problem. It may be interpreted in a number of ways (Hajek 2007), and as such, a number of different potential solutions have been

offered (e.g., Strevens 2016; Wallman and Williamson 2017). It has a long and distinguished history in the field of statistical analysis, beginning with Venn (1866), and it was framed in what is now the standard way by Reichenbach: "If we are asked to find the probability holding for an individual future event, we must first incorporate the case in a suitable reference class. An individual thing or event may be incorporated in many reference classes, from which different probabilities will result. This ambiguity has been called the problem of the reference class" (Reichenbach 1949, 374).

When inferring the probability of a heart attack in an individual, we may know that she belongs to a reference class in which the probability is 3% (e.g., the population quoted in Fuller's article), but she may also have diabetes, be a smoker, or be an athlete. As a member of these other reference classes, we may suspect that her propensity for having a heart attack is higher or lower than 3%, but what is it? When a reference class problem occurs, one cannot rationally claim that the individual propensity is equal to the AR from any of the relevant reference classes. If one needs to act, one must have a credence to act on, and this credence must somehow take into account all the available evidence and so all the relevant reference classes. The reference class problem is therefore a problem for formulating credences in the face of multiple reference classes.

Consider the fact that the therapeutic effects of two drugs may sum, cancel each other out, or result in a negative effect (via side effects). We may know the effect of each drug independently and have reference classes to draw on to estimate the probability of an effect. However, what if an individual takes both drugs? If we do not know how the two causes interact and we know both are relevant to the outcome, there is no way of rationally formulating a credence in the outcome for that individual. As a minimum, then, we must have some way of estimating the interaction between the causal variables defining the different reference classes. If that evidence is not available, then there is no credence to rationally adopt because there is no procedure for inferring a probability.

If we do not know the propensity in the individual case, and we have multiple reference class ARs to aggregate, what should the rational credence in the outcome in the single case be? Simply interpreting ARs as credences in the individual patient does not tackle this problem at all. The reference class problem is a problem for forming credences because of the difficulties in formulating individual probabilities. More importantly, it is this problem that gives rise to the worry that the individual propensity is not equal to the AR in a population in the first place. Fuller is right to highlight this worry, but he completely fails to identify its source.

4. The reference class problem and individual propensities

If the real problem of applying epidemiological evidence in the individual case is the reference class problem, then Alan Hajek (2007) argues that a subjective interpretation of probability must either be radically permissive but vacuous or be rationally constrained but encounter the reference class problem.

Fuller claims that "the epistemic interpretation has the additional benefit of restoring the importance of clinical judgment and expertise as well as the epistemic function of evidence to our understanding of epidemiological evidence in medicine"

(Fuller 2020, 1128). He may be right that inferring individual case propensities demands some kind of judgment, but he avoids mentioning what this could possibly be. More importantly, simply taking evidence from a number of reference classes and producing a subjective credence using some vague notion of "clinical judgment" risks allowing unconstrained (or poorly constrained) credences to inform clinical decisions. In order to avoid this, "clinical judgment" must represent some kind of process that includes rationally constraining credences based on objective evidence. However, it is easy to see how the reference class problem threatens any rationally constrained subjective theory (and for a thorough examination, see Hajek [2007]).

Now, I do not think that the reference class problem makes single-case credences necessarily irrational. If we were not able to form reasonably accurate credences in single cases based on multiple sources of information, we would not be able to safely step out the front door, let alone make complex decisions in medicine. The point I make here is that for probability to be a guide to life, we must first estimate the propensity of chance setups to produce certain outcomes *and then calibrate our credences to these propensities.* The reason evidence ought to constrain our credences is that the evidence informs us about propensities, and propensities are what our credences are meant to track. How to accurately estimate single-case propensities, and why we sometimes fail to do so, is where all the interesting problems and difficulties lie. Formulating a credence once we have estimated the likely single-case propensity is not a problem at all, and recommending that we do this does not solve any real issue.

Distinguishing the theoretical notion of a single-case probability that is meant to represent propensities and the credences that are meant to track them clarifies what is at stake. Propensities are the result of particular generating conditions of chance setups and will vary as those conditions vary. It is for this reason that Fuller correctly argues that individuals in a population may vary in their dispositional profiles, and therefore their propensities, from each other and the populations they are members of. One general approach that tackles this problem, as previously mentioned, is given by Mauricio Suárez (2018), who provides an account of propensities as indexed probability functions. By clarifying the dependence of propensities on their generating conditions, Suárez provides a theory of propensity suitable for the individual case where the "relationship between such propensities and the chances that they bring about is thus to be understood on the model of the 'manifestation' relation in the metaphysics of dispositions" (Suárez 2018, 1156).

Given the dependence on generating conditions, propensities are also thought to be causal in nature, and many of the solutions offered for inferring single-case propensities seek to derive them from causal models rather than aggregates of reference classes. It is not possible to do justice to the variety of options on offer, but two notable approaches include Strevens's (2016) approach for inferring propensities from the causal properties of a setup and Wallmann and Williamson's (2017) methods for setting credences in accordance with causal models. The point to make here is simply that the issue for inferring propensities in the individual case is the problem of appropriate causal or statistical modeling. It is not a question of how we interpret the meaning of frequencies in populations. Once we formulate a causal model, we can then infer the individual propensity in question, and rationality then dictates that our credence should match it.

The role of evidence from AR and frequencies in populations is then seen as informing casual modeling, which is then applied in particular cases.

5. Conclusion

In this article, I start by clarifying the relationship between probability, propensities, and credences, following Humphreys's (1989, 2004, 2019) approach. It is then clear that although Fuller (2020) correctly identifies that there is a problem with using epidemiological studies to quantify the probability of an outcome in an individual case, in recommending that we interpret probabilities as credences, Fuller misidentifies the issue and so fails to offer a meaningful solution to it.

The problem that arises when we try to infer what will happen to an individual based on frequency data in a population is the problem of inferring the propensity for a particular event occurring that involves that individual. When there are multiple reference classes to which that individual belongs, rationality dictates that we take all of them into account when inferring the individual's propensity, and this leads to the reference class problem. There are numerous approaches to solving this problem, which are beyond the scope of this article, but Fuller's recommendation that we interpret probabilities as credences in the individual case is not especially helpful.

References

Beebee, Helen, and David Papineau. 1997. "Probability as a Guide to Life." *Journal of Philosophy* 94 (5):217.

Fuller, Jonathan. 2020. "Epidemiological Evidence: Use at Your 'Own Risk'?" *Philosophy of Science* 87 (5):1119–29.

Gillies, Donald. 2000. *Philosophical Theories of Probability.* London: Routledge.

Goff, David C., Jr., Donald M. Lloyd-Jones, Glen Bennett, Sean Coady, Ralph B. D'Agostino, Raymond Gibbons, Philip Greenland, Daniel T. Lackland, Daniel Levy, Christopher J. O'Donnell, Jennifer G. Robinson, J. Sanford Schwartz, Susan T. Shero, Sidney C. Smith, Jr., Paul Sorlie, Neil J. Stone, and Peter W. F. Wilson. 2014. "2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines." *Circulation* 129 (25, Suppl. 2):S49–S73.

Hajek, Alan. 2007. "The Reference Class Problem Is Your Problem Too." *Synthese* 156 (3):563–85.

Humphreys, Paul. 1989. *The Chances of Explanation: Causal Explanation in the Social, Medical and Physical Sciences.* Princeton, NJ: Princeton University Press.

Humphreys, Paul. 2004. "Some Considerations on Conditional Chances." *British Journal for the Philosophy of Science* 55 (4):667–80.

Humphreys, Paul. 2019. *Probability and Propensities. Philosophical Papers.* Oxford: Oxford University Press.

Lewis, David. 1980. "A Subjectivist's Guide to Objective Chance." In *Studies in Inductive Logic and Probability*, vol. 2, edited by Richard C. Jeffrey, 267–97. Berkeley: University of California Press.

Lyon, Aidan. 2014. "Rényi: There's No Escaping Humphreys' Paradox (When Generalized)." In *Chance and Temporal Asymmetry*, edited by Alastair Wilson, 112–25. Oxford: Oxford University Press.

Reichenbach, Hans. 1949. *The Theory of Probability.* Berkeley: University of California Press.

Strevens, Michael. 2016. "The Reference Class Problem in Evolutionary Biology: Distinguishing Selection from Drift." In *Chance in Evolution*, edited by C. Pence and G. Ramsey, 145–75. Chicago: University of Chicago Press.

Suárez, Mauricio. 2018. "The Chances of Propensities." *British Journal for the Philosophy of Science* 69 (4):1155–77.

Venn, John. 1866. *The Logic of Chance: An Essay on the Foundations and Province of the Theory of Probability, with Especial Reference to Its Logical Bearings and Its Application to Moral and Social Science.* London: Macmillan.

Wallmann, Christian, and Jon Williamson. 2017. "Four Approaches to the Reference Class Problem." In *Making It Formally Explicit: Probability, Causality and Indeterminism*, edited by Gabor Hofer-Szabo and Leszek Wronski, 61–81. Cham, Switzerland: Springer.