Synthese (2021) 199:12251–12269
https://doi.org/10.1007/s11229-021-03331-0

ORIGINAL RESEARCH

When statistical evidence is not specific enough

Marcello Di Bello[1]

Received: 7 December 2020 / Accepted: 21 July 2021 / Published online: 12 August 2021

Abstract
Many philosophers have pointed out that statistical evidence, or at least some forms of it, lack desirable epistemic or non-epistemic properties, and that this should make us wary of litigations in which the case against the defendant rests in whole or in part on statistical evidence. Others have responded that such broad reservations about statistical evidence are overly restrictive since appellate courts have expressed nuanced views about statistical evidence. In an effort to clarify and reconcile, I put forward an interpretive analysis of why statistical evidence should raise concerns in some cases but not others. I argue that when there is a mismatch between the specificity of the evidence and the expected specificity of the accusation, statistical evidence—as any other kind of evidence—should be considered insufficient to sustain a conviction. I rely on different stylized court cases to illustrate the explanatory power of this analysis.

Keywords Naked statistical evidence · Criminal trials · Eyewitness testimony · Probability · Decision · Standards of proof · Specificity

1 Introduction

The literature on naked statistical evidence spans several decades and has seen a recent resurgence. Philosophers and legal theorists have pointed out that statistical evidence lacks desirable epistemic or non-epistemic properties, and that this should make us wary of litigations in which the case against the defendant rests in whole or in part on statistical evidence.[1] Other authors have responded that reservations about statistical

[1] See, for example, Tribe (1971), Nesson (1979), Cohen (1977), Thomson (1986), Wasserman (1991), Stein (2005), Ho (2008), Enoch et al. (2012), Cheng (2013), Buchak (2014), Pritchard (2015), Blome-Tillmann (2015), Nunn (2015), Staffel (2016), Smith (2018), Pundik (2017), Littlejohn (2020), Gardiner (2018), Di Bello (2019), Dahlman (2020), Bolinger (2020), Moss (2021), Nelkin (forthcoming).

This article belongs to the topical collection "Recent Issues in Philosophy of Statistics: Evidence, Testing, and Applications", edited by Sorin Bangu, Emiliano Ippoliti, and Marianna Antonutti.

✉ Marcello Di Bello
mdibello@asu.edu

[1] School of Historical, Philosophical and Religious Studies, Arizona State University, Tempe, AZ, USA

evidence are misplaced because appellate courts have expressed nuanced views.[2] In DNA evidence cases, for example, appellate courts in the United States have upheld convictions that crucially rely on base rate statistics about the frequency of genetic profiles.[3]

No doubt statistical evidence comes in different forms, and some may raise concerns more clearly than others. In an effort to clarify and reconcile, I will examine different forms of statistical evidence as they are used in criminal trials. For the sake of readability, this examination will rely on stylized court cases and is not intended to be exhaustive. Concomitantly, I will put forward a theory of whether—and if so, under what circumstances—statistical evidence should raise opposition when it is used to establish the defendant's guilt. The key idea in my argument is that of specificity. I argue that when there is a mismatch between the specificity of the evidence and the expected specificity of the accusation, statistical evidence—as any other kind of evidence—should be considered insufficient to sustain a conviction unless it is adequately supplemented by other, more specific evidence.[4] This analysis explains why we intuitively resist a conviction in hypothetical cases of naked statistical evidence, even though our intuitive resistance is less pronounced in other cases, such as DNA evidence cases. At the same time, this analysis is inevitably incomplete. As the argument of the paper will make clear, the degree of specificity that a criminal accusation should possess cannot be defined apriori and varies on a case-by-case basis. This variability might well drive the disagreement among philosophers and legal scholars about the proper place of statistical evidence in criminal trials.

Two limitations of my analysis are worth mentioning at the outset. First, the discussion that follows is confined to criminal trials. Towards the end, I offer some thoughts about statistical evidence in civil trials, but these are merely tentative. Since civil and criminal trials differ in important ways—the stakes are higher in criminal than civil trials; the standards of proof are different; etc.—they are best examined separately. Second, my concern is with the question of sufficiency, not admissibility.[5] There may be reasons not to admit statistical evidence at trial even when it only plays a supplementary role.[6] Concerns about admissibility aside, the paper examines cases in which statistical evidence plays a key role in the prosecutor's argument against the defendant. The paper asks whether, in such cases, statistical evidence suffices to sustain a conviction.

A few words on terminology are also in order. The literature often speaks of 'naked' statistics—meaning statistical evidence unaccompanied by other evidence. This is too strong as there is inevitably going to be other evidence. As the name suggests, naked statistical evidence consists of statistics, say a base rate frequency or a quantity derived from a probability model (such as the expected frequency of a genetic profile in a population). Naked statistical evidence is 'naked' not so much in the sense that it is

[2] See, for example, Kaye (1979), Schmalbeck (1986), Schoeman (1987), Allen and Leiter (2001), Schauer (2003), Redmayne (2015), Hedden and Colyvan (2019), Krauss (2020), Ross (2021), Papineau (2021).

[3] For a representative set of cases, see Malcom (2008), Roth (2010).

[4] For an earlier versions of this specificity argument, see Chapter 7 of Di Bello (2013).

[5] On the distinction, see the helpful remarks by Picinali (2016).

[6] See, for example, the discussion in Di Bello and O'Neil (2020). For a helpful review of the case law in the United States on the question of admissibility and statistical evidence, see Koehler (2002).

the only evidence in a case—there is always other evidence—but in the sense that it plays a crucial role in the prosecutor's argument against the defendant. That is, had it not been for the statistics, the prosecutor would have no chance to prevail. This characterization is inevitably vague, and unfortunately, the lack of a clear definition is endemic to the literature. I will, however, provide several examples which should partly remedy the lack of a clear definition.

I begin with a paradigmatic example of naked statistical evidence in criminal trials (Sect. 2). I then show how the notion of specificity helps to understand and justify people's resistance toward a conviction (Sect. 3). The rest of the paper refines the specificity argument in a number of contrast cases: eyewitness testimony (Sect. 4); other cases of naked statistics (Sect. 5); cases of statistics that are unproblematic (Sect. 6); statistics-based identification evidence (Sect. 7) including cold-hit DNA matches (Sect. 8). The paper concludes (Sect. 9) with a few remarks about the relationship between probability, specificity and the availability of evidence at trial.

2 Workers

Imagine a factory where 100 workers are assembling pieces of equipment, while a guard supervises them. Suddenly 99 of them walk away from their work stations, approach the guard and kill him. Surveillance cameras show a few workers holding the guard, others taking turns in punching him until the guard falls on the ground. The video recording shows that one worker refrained. The others pushed him against the guard but he clearly did not want to have anything to do with the incident. We cannot tell who he was. The workers were identically dressed, and the video recording was of poor quality. There were no other witnesses. The video was the only evidence.

After the fact, the police promptly arrive at the scene and find everybody back at work. They see the guard's dead body and quickly realize that something is not right. After watching the surveillance video, the police decide to pick one worker at random. Since 99 out of the 100 workers present killed the guard, there is a 99% chance that the worker picked at random is one of the culprits. The police believe that is enough to arrest him. The worker is then charged with manslaughter, assault and rioting. At trial, the prosecutor argues that the evidence against the worker is overwhelming. It is extremely likely that he was one of the perpetrators. The worker is convicted.

Scenarios such as this one have been around for a while.[7] For those willing to engage with such scenarios, a common reaction is uneasiness. True, it is very likely that the worker is guilty. But the high probability of guilt does not seem enough to justify a conviction. The descriptive puzzle here is, what is the cognitive mechanism that explains why we resist a conviction even though the probability of guilt is so high? The normative puzzle is, what is the justification for why we should resist a conviction even though the probability of guilt is so high? In what follows, I will be concerned with the normative puzzle.[8]

[7] For variations of this scenario, see Nesson (1979) and Cohen (1977).

[8] Psychologists and experimental philosophers have examined the descriptive puzzle, sometimes called the Wells' effect; see, for example, Wells (1992), Niedermeier et al. (1999), Ebert et al. (2018).

As a preliminary step, it is important to reflect on the scenario itself. Legal practitioners will observe that the worker scenario is far-fetched. It is odd that the worker was picked at random and brought to trial without conducting an investigation. The police should arrest and interrogate every worker, as well as examine the crime scene more carefully. If they did, the evidence would not just consist of the naked statistics—that is, the numerical fact that 99 workers out of the 100 workers present killed the guard. There would be other evidence. This other evidence could show that the worker on trial was one of the killers or alternatively that he refrained from killing the guard. The scenario would then self-destruct and the puzzle would evaporate. I call this the *practitioner's stance*.[9]

The practitioner's stance offers a first pass at a solution. Perhaps, a conviction should be resisted in the worker scenario simply because it would be based on a body of evidence whose scope is artificially narrow.[10] But an obvious response—from those more willing to engage with hypotheticals—is that the worker scenario tacitly assumes that no other evidence could be found. The scenario assumes there were no eyewitnesses who could help identify the perpetrators; all the workers were unwilling to talk; etc. The puzzle then recurs: assuming no further evidence could be found, why should we resist convicting the worker on trial if the odds of his guilt are so overwhelming?

A natural way to answer this question is to focus on the statistical nature of the evidence. The evidence consists in the numerical fact that 99 workers killed the guard and one did not. This numerical fact provides information about a group, not about an individual. As a consequence, given the 99-to-1 statistics, any worker picked at random is 99% likely to be one of the perpetrators and 1% likely to be the innocent one who stood apart from the others. The statistics fail to single out the worker on trial and distinguish him from the others. If the evidence presented at trial should be tied specifically to the defendant or the events under dispute, the 99-to-1 statistics do not satisfy this requirement and thus cannot serve as evidence about what happened.[11]

[9] For a critique of the use of hypothetical scenarios in theorizing about evidence law, see Allen (2021).

[10] Some scholars speak of *reasonable completeness* of the evidence or lack thereof. Kaye (1986) gives an anecdotal example of a drunk driving case in which the prosecutor called the arresting officer to the stand who testified about the smell and breath of the man arrested, as well as the results of a field sobriety test. The incriminating evidence did not, however, include the results of a breathalyzer. Kaye notes that the jurors expected to hear about that type of evidence but did not, and thus acquitted the defendant. A similar idea is invoked by Judge Richard Posner in Howard v. Wal-Mart Stores Inc. 160 F.3d 358 (7th. Circ. 1998). Nance (2016) in a recent book gave a more detailed account of what it means for the evidence to be reasonably complete—that is, it should be all the evidence that someone tasked with making a decision at trial would reasonably expect to see from a conscientious investigation of the facts in the type of case at hand.

[11] The literature has refined this point in different ways, drawing on epistemological and moral considerations. According to Thomson (1986), the 99-to-1 statistics are not causally connected, in the appropriate manner, with the facts of the case, while other forms of evidence, say eyewitness testimony, typically are. Colyvan et al. (2001) and Allen and Pardo (2007) emphasize the reference class problem which tend to affect inferences made on the basis of quantitative information. Enoch et al. (2012) and Pritchard (2015) argue that naked statistics are not specific because they lack the modal properties of sensitivity and safety. Wasserman (1991) and Pundik (2017) point out that convictions are attributions of individual (as opposed to group-based) culpability, and naked statistical evidence conflicts with this assumption since it describes group-level features.

And yet, the 99-to-1 statistics make it 1% likely that the worker on trial is innocent. Shouldn't such a low probability be enough to rule out the hypothesis of innocence? People cannot be convicted of a crime unless their guilt is established beyond a reasonable doubt. But, some will argue, since the expression 'beyond a reasonable doubt' does not mean 'absence of any doubt whatsoever', a 1% risk of error seems low enough to meet the bar.[12] Others will respond that 'proof beyond reasonable doubt' cannot be understood as the requirement that guilt be established with sufficiently high probability and innocence be ruled out with sufficiently low probability.[13] That proof beyond a reasonable doubt cannot be understood as a probability threshold is—some will insist—the key lesson to learn from cases of naked statistical evidence such as the worker scenario.

The discussion so far highlighted three strategies for addressing the puzzle of naked statistical evidence. The first strategy resists the scenario itself by pointing out that the evidence is artificially narrow in scope. The second strategy focuses on the statistical nature of the evidence and identifies its deficiencies, say the statistics are not individually tied to the defendant on trial. The third strategy is more general and consists in rejecting a simple probability threshold interpretation of the standard of proof in criminal cases. These strategies are intertwined in important ways, and I believe that each has merit. But instead of defending one strategy or the other, I want to focus on an issue that has not received much attention in the literature, namely how detailed the accusation of guilt should be. The problem in the worker case—I hold—is not so much that guilt is framed in probabilistic terms or that the evidence is statistical, but rather, that we should not think of guilt—or more precisely, the accusation of guilt—in a generic and unspecified way. Towards the end of the paper, I will comment on how my line of argument overlaps with the other three strategies for tackling the puzzle of naked statistical evidence.

3 Specificity

The central idea in my argument is that of specificity. The statistics in the worker case say little about what happened. Even assuming no other evidence could be found, the 99-to-1 statistics are still weak evidence of guilt because they provide little information about what the worker on trial did. Was he one of the people holding the guard? Was he one of those punching the guard? These questions are left with no answer.[14] The recording shows with reasonable precision what happened and what each worker did. But it is impossible from the video to tell who did what. As fact-finders in a trial, we need not worry too much about what other workers did, but we need to know what the worker on trial did.

The lack of specificity makes the evidence against the worker a thin ground to justify a conviction. This is not a shortcoming to the 99-to-1 statistics *qua* statistics.

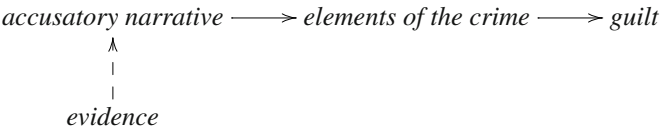
[12] The debate in the literature about the meaning of 'proof beyond a reasonable doubt' is ongoing. For helpful discussions, see among others Weinstein and Dewsbury (2006), Laudan (2011), Gardiner (2019).

[13] See, for example, Stein (2005), Ho (2008), Allen (2010), Haack (2014), Gardiner (2019), Moss (2021).

[14] Schauer (2021) makes a similar point and draws a connection to the so-called problem of aggregation as discussed in Porat and Posner (2012).

Non-statistical evidence could very well exhibit the same lack of specificity (more on this later). Nor is the lack of specificity a property of the evidence itself. Rather, it is a property of the inferences that can reasonably be drawn from the evidence. To make this point more precise, a few words about how evidence is deployed at trial are in order.

When a defendant is charged with a crime and later faces trial, the prosecutor is tasked with establishing the defendant's guilt in a manner commensurate to the charge (murder, rape, theft, etc.). The prosecutor should prove the elements of the crime from which guilt follows in accordance with the substantive law governing the case. To establish the elements of the crime, the prosecutor usually puts forward an accusatory narrative (story, theory) about what happened.[15] For instance, in homicide cases the narrative of the crime should contain information concerning who committed the crime, when, where, how, and also why it was committed. Of course, the prosecutor does not merely put forward an accusatory narrative. The latter should also be adequately supported by the evidence, say it should fit with the evidence or be highly probable given the evidence.[16] The relation of evidence, narrative, elements of the crime, and guilt can be visualized as follows:



The accusatory narrative consists of the inferences that the prosecutor draws—and would like others to draw—from the evidence. Narratives can be more or less specific. A narrative asserting that the perpetrator was a male with fair complexion in his fifties is more specific than a narrative describing the perpetrator as an adult male; a narrative which offers information about the identity of the perpetrator as well as the time and place of the crime is more specific than a narrative which only offers information about the perpetrator's identity; and so on. The specificity of a narrative can also be defined in terms of *informativeness*—i.e. the extent to which it can answer all the relevant questions in a case.[17]

To see how all this applies to the worker scenario, it is helpful to examine the possible charges that could be leveled against the worker and see if the 99-to-1 statistics can support a narrative that would sustain those charges. Start with manslaughter. The charge of manslaughter is the allegation that the worker intentionally killed the guard, albeit without premeditation or malice aforethought. This charge would be difficult to sustain in court on the basis of just the 99-to-1 statistics. Is the worker on trial one of those who caused the guard's death? Could the worker have injured the guard without causing his death? The statistics cannot address these questions. The prosecutor could

[15] I am indebted to the 'story model' of judicial decision-making; see Pennington and Hastie (1991), Urbaniak (2018), van Koppen and Mackor (2020). See, however, Allen and Pardo (2019) who prefer the notion of 'explanation.'

[16] Fitelson (2006) offers an excellent discussion of many probability-based accounts of evidential support.

[17] On informativeness, see Carnap and Bar-Hillel (1952), Groenendijk and Stokhof (1997) and Floridi (2004).

pursue a less serious charge, such as assault, usually defined as the harmful or offensive contact with another person without that person's consent. Another possible charge is one of rioting, sometimes defined as the use of force or violence by two or more people acting together. From the video recording, it is clear that all 99 workers, in some way or another, physically injured the guard. But again, the recording does not provide information about the extent to which each worker—including the one on trial—caused harm to the guard.

So what kind of accusatory narrative do the 99-to-1 statistics actually support? They support a rather unspecified narrative of what happened, along the following lines: the worker on trial, together with ninety-eight unidentified other workers, attacked the guard and killed him *by doing something or other*. This narrative is 99% likely on the evidence and entails the proposition 'The worker participated in the crime' from which guilt follows. But what did the worker on trial do during the killing? What role did he play? How did he kill the guard? To all these questions, the unspecific narrative gives no answer.

In contrast, consider the more specific accusatory narrative that, say, the worker on trial held the guard at first with the help of three other workers, urged other workers to punch the guard and then punched the guard several times. This more specific narrative is not very likely to be true given just the 99-to-1 statistics. So, if the prosecutor had offered the more specific narrative, it would be clear that the 99-to-1 statistics did not support it.

The point here is that prosecutors do not establish someone's guilt in a vacuum. They establish someone's guilt by offering an accusatory narrative about what happened and providing evidence that supports it. The choice of the accusatory narrative is not arbitrary. What the criminal justice system owes each defendant, especially in homicide cases, is a well-specified accusation—one that describes what they did in a time and place and that can be subsumed under a typology of crime that is punished by law.[18] People are not typically accused of having committed a crime in some way or another. So, in the worker case, there is a mismatch between the unspecified information the 99-to-1 statistics can provide and the expectation that a conviction should detail what the accused did.

[18] For more on this point, see Duff (2001), Burns (2004). That criminal accusations should usually be specific is a peculiarity of criminal trials. There are many situations in which the same degree of detail is not required. Consider a bank that gives out loans. The bank uses a scoring system that draws on factors such as assets, credit history and job security. Risk scores track the likelihood of default if a loan is granted. Scores range between 0 (zero likelihood of default) and 100 (certain default). Suppose Tera needs a loan to purchase a home. She applies for a loan and is assigned a risk score of 3. The bank deems her at low enough risk of default and grants her the loan. The bank can do business without settling the question whether Tera—specifically—will default or not. If the scoring system is trustworthy, Tera's default is 3% likely. This is all the bank needs to know. Since the bank expects that 3 out of 100 people with a risk score of 3 will default, it will adjust the interest rate appropriately. So banks—just like casinos and insurance companies—do not need individualized information. They make decisions about long-run patterns. So they can routinely rely on the relevant statistical or actuarial information without problem.

4 Eyewitness testimony

I mentioned earlier that the lack of specificity is not confined to statistical evidence. As I show in this section, non-statistical evidence can also fail to be specific in the required manner. Conversely, statistical evidence can be specific in the required manner, say in cases in which the accusation is about a pattern of behavior (more on this later in the paper). So the question of specificity is orthogonal to the statistical nature of the evidence.

Consider this example. A robbery occurred in New York City on 5th Avenue on January 20, 2020. An eyewitness testifies 'I saw the defendant ran along 5th Avenue on January 20, 2020'. This testimony provides information about the defendant's whereabouts in a specific time and place. And yet, the testimony is not sufficiently specific. The testimony places the defendant in the vicinity of the crime scene. It addresses the question 'Was the defendant there?' If the defendant was in the vicinity of the scene, they had an opportunity to commit the crime. That is useful information, but does not address other relevant questions. How did the defendant carry out the robbery? Were there any accomplishes? Etc. The testimony—even if believed to be trustworthy—is clearly insufficient for a conviction. It must be supplemented by other evidence. So any evidence, whether it is statistical or testimonial, can fail the specificity test. Specificity depends on whether the evidence addresses the questions that the criminal conviction is supposed to address. A testimony that says 'I saw the defendant run away from the crime scene' is not as specific as 'I saw the defendant's face while he was stabbing the victim in the chest.'

There is another, less compelling reason why eyewitness testimony could lack specificity. Suppose a psychologist testifies that in normal circumstances eyewitnesses misidentify people in 30 out of 100 cases, and in stressful circumstances, the error rate raises to 50 out of 100.[19] Here the 30-to-100 or 50-to-100 error statistics are supplementary information to assess the trustworthiness of the testimony. But these statistics are not specific to the testimony itself. They provide information about the trustworthiness of eyewitness testimonies *in general*. They are long-run statistics. In this sense, it is tempting to dismiss them as irrelevant because at trial the goal is not to assess the trustworthiness of eyewitness testimonies in general. The goal is to assess the trustworthiness of this testimony.

This worry can be alleviated by noting that no trial testimony should be believed unless it has passed a rigorous adversarial test. In particular, cross-examination is the process by which more information is elicited to assess the reliability of a specific testimony. Here are questions that witnesses may face. How far were you? For how long did you witness the incident? Were you forced to testify? Etc. These case-specific questions, however, are still informed by generalizations of some kind, say that people who are too far cannot usually see clearly. Reliance on these generalizations is inevitable. But such reliance should not be the reason why eyewitness testimony cannot be enough to sustain a criminal charge.

Two questions should be asked in evaluating any piece of evidence at trial. The first question is: How specific is this evidence? That is, how specific is the accusation

[19] On the reliability of eyewitness testimony, see Loftus (1996), Wixted and Wells (2017).

(story, narrative, theory of the crime) which this evidence can support? Specificity is a property of the accusatory narrative, not directly of the evidence itself, but the evidence can be more or less specific to the extent that it can support a more or less specific narrative. The second question is: How reliable and trustworthy is this evidence? In assessing trustworthiness, it is often inevitable to rely on generalizations of some kind or long-run error statistics.

So an eyewitness testimony whose informational value can be more or less specific should be distinct from the long-run error statistics that serve to gauge its trustworthiness. An eyewitness testimony fails the specificity test when the information it provides does not address all the relevant questions, not because long-run error statistics may be used to evaluate its credibility.[20]

5 Person-specificity and event-specificity

In the worker scenario, the overwhelming probability of guilt is not enough to justify a conviction because—I argued—the 99-to-1 statistics are not specific enough. But when is the evidence specific enough? A simple way to flesh out their lack of specificity is to point out that the 99-to-1 statistics apply indiscriminately to all the members of the group of workers during a particular shift and cannot single out the one innocent worker. On this reading, the statistics lack specificity because they cannot single out the worker on trial from the others. I call this person-specificity.

While helpful in its own way, person-specificity does not offer a complete account of specificity. In fact, it would be incorrect to think that whenever the evidence can distinguish between one suspect and others—say, it makes one suspect extremely more likely to be guilty than others—the evidence counts as sufficiently specific. Things are not so simple.

Suppose two individuals, Feira and Viara, are browsing through the clothes in a department store early in the morning, right after the store has opened. There are no other customers in the store. The two individuals later purchase two moderately priced items. At the cashier desk, they start talking to one another and end up leaving the store together. As they exit the store, the alarm goes off. The security guard approaches them and conducts a search through their bags, but finds no stolen merchandise. Feira and Viara are let go. Soon thereafter, a member of the staff finds that an expensive dress is missing from the racks. In all likelihood, the culprit is one of the two customers who just left.

[20] The distinction between the informativeness of the testimony and the error-statistics used to qualify its reliability helps to see why the statistics in the worker scenario are not specific enough. Suppose the first-order evidence in the worker case is the statement 'I saw the worker in the factory where the guard was killed'. Further, suppose the 99-to-1 statistics are used to qualify this testimony. They offer the additional information that 99 of the 100 workers in the factory killed the guard. For a similar analysis of naked statistical evidence, see Dahlman (2020). This formulation makes clear that the first-order information bearing on the incident—namely the testimony that the worker was in the factory—cannot be specific enough to sustain a homicide accusation. Adding the 99-to-1 statistics does not change this fact. Likewise, adding error statistics does not change the fact that the assertion 'I saw the defendant ran along 5th Avenue' is not specific enough to sustain, say, a robbery charge.

The two customers, Feira and Viara, are tracked down and arrested, but each denies having stolen anything from the store. There is no evidence against them except a bit of character evidence paired with statistics. Feira has scored high in a compulsivity test and has been arrested for stealing in department stores several times in the past. Viara, instead, has never been arrested for stealing in a department store and shows no sign of high compulsivity. Statistics show that people with a high degree of compulsivity and who have stolen merchandise in department stores before are more likely than others to steal merchandise if they are unsupervised. So Feira is most likely the culprit.[21] For the sake of illustration, suppose studies show that people like Feira, when unsupervised, will steal 99 times out of 100 times. Instead, people like Viara, when unsupervised, will only steal 1 time out of 100 times. So Feira is 99 times more likely than Viara to have stolen the merchandise. Would these statistics be strong enough evidence to convict Feira?

As in the worker case, many would feel uneasy about convicting (or even arresting) Feira on such thin evidence. No merchandise was found on her. She could not be directly linked to the crime. The notion of specificity can help here. Feira is accused of having stolen merchandise in a particular department store at a particular time. But the statistics only show that Feira would steal at some department store at some point—in fact 99 out of 100 department stores. They do not provide information that specifies why this time is any different from another time. There is a mismatch between the information the statistics can provide and the degree of specification that is expected to sustain the accusation.

The statistics in this case are specific to some extent, however. They make a difference between Feira and Viara. They show that Feira, unlike Viara, has stolen or will steal from department stores, if unsupervised, a large number of times. But the statistics still make no difference between what Feira would do this time as opposed to another time she visits a department store. How is this visit different from any other? In that sense, the statistics against Feira are not specific enough. They make it extremely likely that Feira would steal at *any* department store. But they do not say that it is extremely likely that she would steal at this particular department store as opposed to another. The statistics are useful to establish a pattern of behavior, not a behavior in a specific time and place.

In the department store example, the statistics are person-specific. Their lack of specificity is not due to the fact that they cannot uniquely single out the perpetrator. They can single out one suspect, Feira, as more likely than another, Viara, to have stolen the merchandise. But although the statistics are person-specific, they are not event-specific. They show that Feira, unlike Viara, is likely to steal at any department store. In the worker example, instead, the 99-to-1 statistics are event-specific. They are about what the workers did in that particular moment in time. They would not be applicable to another riot or alike event in the past or the future. We know, in that time and place, that 99-to-1 workers assaulted and killed their supervisor. What the statistics are missing, though, is person-specificity.

[21] For similar examples, see Pundik (2017) and Gardiner (2020).

6 Expected specificity, not plain specificity

The discussion so far highlighted different forms of specificity, person-specificity and event-specificity, but has not yet answered the crucial question. When is the incriminating evidence specific enough? Perhaps, the incriminating evidence counts as specific enough if it is both person-specific and event-specific. But event-specificity—and to some extent person-specificity—come in degrees. There is no limit to how specific the description of an event could be. For suppose there is overwhelming evidence against a defendant. The evidence shows in great detail how the defendant committed a bank robbery: time, place, strategy, accomplices, etc. The evidence satisfies event-specificity and person-specificity to some degree. But there are gaps in the narrative here and there. For example, it is unclear whether the defendant entered the bank from an underground tunnel on the north or south side of the building. Would such lack of event-specificity be enough to block a conviction? At first blush, the location of the tunnel would seem an irrelevant detail in light of the mounting evidence against the defendant. But the matter is not so clear cut.

Recall that my claim is that convictions cannot be solely based on statistical evidence—in fact, on any other evidence—whenever there is a mismatch between the degree of specificity of the accusation which the evidence can support and the degree of specificity that is *expected* to sustain the charges. But how specific the accusatory narrative (or simply the accusation) should be cannot be defined apriori. An accusation is specific enough when it is informative enough to address all the relevant questions in a case.

Different factors are at play here. One factor for assessing how specific an accusation should be is the negotiation of the parties. Does the prosecutor need to establish what the defendant did at every point in time? Surely not, but certain details will make a difference; others will not. Return to the worker example. Whether the worker on trial was one of those who attacked the guard or the lone worker who stood in a corner makes a significant difference. It is the difference between guilt and innocence. So providing information that helps to answer questions about the worker's whereabouts is crucial. The 99-to-1 statistics are unhelpful in this respect and thus are a weak basis for a conviction.

Or consider the bank robbery case above. Does the prosecutor need to establish which underground tunnel the robber used? In many cases, this would be an irrelevant detail given the strong evidence of guilt against the defendant. But suppose the defense lawyer argued at trial that the defendant must be innocent because (a) the robber could only use the north tunnel and (b) the defendant could not have used the north tunnel since the location of the north tunnel is incompatible with the defendant's whereabouts. If so, a seemingly small detail would become crucial.

Another factor that affects how specific an accusation should be is the type of crime being prosecuted and its circumstances. Criminal charges are often—not always—about a crime committed in a specific time and place. Accusations that lack specificity relative to space or time are typically suspect. This is often true in homicide accusations. But accusations can also be less specific. This depends on how crimes are defined and prosecuted. As we will now see, when accusations describe patterns of behavior over time, the requirement of event-specificity is relaxed. A shift then occurs

about what needs to be proven, a shift from showing causation of harm in this instance to mere exposure to a heightened risk of harm over a period of time. This shift is likely to render statistics about patterns of behavior more palatable as incriminating evidence.

I am going to give two illustrative examples of this phenomenon. First, suppose a nurse is on duty at the hospital during certain days of the week. When this nurse is on duty, an unusually large number of patients die. The number is much higher than when other nurses are on duty. The hospital director suspects the nurse was negligent and did not provide adequate medical care to the patients. Other plausible explanations are ruled out. Shifts are assigned randomly to the nurses during any day of the week or time of day. Patients are also assigned randomly to the nurses. So the cause of the unusually high number of deaths recorded seem to be the nurse's behavior. There is no specific evidence that the nurse caused the death of this or that patient. But the evidence points to a long-term pattern. This pattern can be used as incriminating evidence provided the accusation does not concern any specific incident. Perhaps the nurse could be accused of exposing their patients to a heightened risk of death compared to other nurses. If that is the accusation, the statistics would be adequate evidence to sustain it.[22]

Consider another example. Suppose a trader orders several financial products, but ends up buying a fraction of them. This is not uncommon, but compared to general market trends, this particular trader finalizes a much smaller percentage of the orders. Suppose also such orders affect the market value of the financial products. So the trader is accused of price manipulation. While there is no specific evidence of what the trader did in any specific occasion, the trader's pattern of buying dramatically deviates from the norm. The anomaly raises suspicion. Such statistical evidence could be used to mount a case against the trader for price manipulation.[23]

The moral here is twofold. First, the required degree of specificity of the accusation may vary on a case-by-case basis. For one thing, even when the evidence satisfies person-specificity and event-specificity at some level of granularity, the evidence may still not be specific enough. The bank robbery example illustrates this point. On the other hand, it may sometimes be too strict to expect that the evidence be both person-specific and event-specific. The accusation need not always be extremely detailed, for example, when it concerns a pattern of actions over an extended period of time. There is no clear-cut formula to determine whether the evidence is specific enough to sustain an accusation. Sufficient specificity depends on policy objectives such as the types of crime being prosecuted, or contextual factors such as which questions become relevant during the adversarial confrontation of prosecution and defense.[24]

The second moral is that, even though long-run statistics are not event-specific, they could still count as specific enough in certain contexts. The nurse and price manipulation examples make clear that statistical evidence is not objectionable per se. It is only objectionable when there is a mismatch between the level of detail expected in the criminal accusation and the degree of detail the statistical evidence can support. If

[22] For court cases along these lines, see Feinberg and Kaye (1991), Meester et al. (2006).

[23] For an analogous case, see US v. Coscia, 866 F. 3d 782 - Court of Appeals, 7th Circuit 2017.

[24] Another factor that may affect the expected degree of specificity of the accusation is the availability of the evidence. It may be hard to reconstruct someone's whereabouts in detail if it is difficult to find adequate evidence. Consider a tort case, leaving the criminal realm for a moment. Say a person has developed cancer

the accusation is not expected to be event-specific, it does not matter that the statistics are not event-specific. In the next section, I am going to consider what happens if the accusation is expected to be event-specific and there is statistical evidence that can meet this expectation. Presumably, since there is no mismatch between the statistics and the accusation as far as the degree of detail required, the statistics should not be problematic. This is precisely what happens.

7 Highly specific statistics

Let me start by considering a futuristic example of statistical evidence that is extremely detailed. This example serves to further buttress the claim that statistical evidence is not objectionable as such. Its lack of specificity is. Suppose extremely detailed statistics—in the form of hundred of correlations elicited through Big Data and Machine Learning—show that Mark Ito will commit a robbery at PNC Bank on 5th Street, Remington on January 16, 2021 at 4:30 PM. The evidence also makes predictions about what the robber will do; how they will conduct the robbery; etc. No doubt this evidence satisfies both person-specificity and event-specificity to a very high degree. We could not have asked for anything more specific than that. Assume this evidence is trustworthy. Similar predictions have been correct 99.99999% of the time. Would that be enough to convict Mark Ito of robbery?

Intuitively no. This is a seeming counterexample to my argument. The statistics are so detailed, and yet they seem insufficient to convict. What is going on here? It is important to distinguish two cases. First, if the robbery has not taken place yet, Mark Ito cannot be preemptively convicted of a crime that has not yet happened, no matter how specific the evidence. Convictions attribute responsibility for what someone did. They are backward-looking. They cannot become forward-looking. However, suppose the robbery has taken place. If the robbery has happened at PNC Bank on 5th Street, Remington on January 10, 2021 at 4:30 PM, this would perfectly match the statistical evidence. In this case, I see no compelling reason why Mark Ito should not be convicted on the statistics alone.

The putative counterexample in fact corroborates the core claim of this paper. That is, people's hesitancy against convicting a defendant on the basis of naked statistical evidence—even when the odds of guilt are high—occurs when there is a mismatch between the level of detail expected in the criminal accusation and the degree of detail the statistical evidence can support. Futuristic evidence that consists entirely of

Footnote 24 continued
while working in a field and using a chemical pesticide. Epidemiological studies show that repeated exposure to the pesticide increases one's likelihood of cancer. Should the company that marketed the pesticide be held responsible? This is a difficult question. But think how hard it would be to hold the company responsible if the requirement was for the plaintiff to show event-specific or individualized causation—i.e. that the pesticide caused the plaintiff's cancer in that particular case. Or consider again the trader example. It would be unreasonable to expect the prosecutor to offer evidence showing what the trader did each time and with what purpose. It would be unreasonable because such evidence would normally not be available. So the degree of specificity of the accusation will also depend on what evidence one would reasonably expect to see given the circumstances of a case.

statistical correlations can well suffice to sustain a conviction when it is sufficiently specific.

The core claim of this paper is also corroborated by more mundane examples of statistical evidence. Trace evidence that heavily relies on statistics—such as DNA matches, shoe prints, glass evidence, etc.—illustrates this point. Consider a burglary case. Say the investigators find certain shoe prints at the crime scene. The prints show where the perpetrator entered; what they took; where they left; etc. The perpetrator broke into the house from a window. They stole TVs, computers and other electronics. The traces leave unanswered one question, though. Who left the prints? Match evidence serves to answer this identification question. If the shoe prints are so rare that only one person could have left them, that would be a full identification. More realistically, experts will testify that, say one person every 500,000 people, on average, would randomly match the prints. If the suspect matches the prints, that is evidence against the suspect. DNA evidence works in the same way. Instead of shoe prints, a genetic profile is created from traces of hair, blood, skin, etc. found at the scene. If the crime scene profile matches a suspect, that is strong evidence against the suspect.

People are less reluctant to base a verdict of criminal liability when matching shoe prints, matching genetic profiles, etc.—more generally, statistics-based match evidence—show that the suspect is very likely to be the perpetrator. The question is whether these forms of statistical evidence differ, in some important way, from the statistical evidence in scenarios such as the worker case or the department store case. Here, the statistics are about the prevalence of a rare identifying trait in a population—shoe print, genetic profile, etc. The identifying trait is rare, say 1/500,000 or even more rare. In the earlier scenarios, the statistics were about the prevalence of a behavior (often closely connected, or even identical to, the guilty behavior). The behavior was widespread in the population, say 99% in the select population are known to engage in that behavior. The puzzle for many in the literature is this. Are both forms of statistics problematic or is there a relevant difference?

I believe that there is a relevant difference here, and this difference can be formulated using the concept of specificity I have been articulating. In the burglary case above, the prosecutor should establish a well-specified narrative of what happened: how the suspect entered the home; what items they stole; etc. Trace evidence—the shape and arrangement of the traces—can say a lot about what happened. But the identification question must be answered, as well. Who left the traces? This question is addressed by the match—for example, a match between the crime scene shoe print and the shoe print of the suspect, or a match between the crime scene genetic profile and the genetic profile of the suspect. The match should be supplemented by quantitive information about how rare the identifying feature is. In this sense, the statistics underlying the match are not problematic as they contribute to addressing the identification question. By contrast, the statistics in the worker scenario—together with the other evidence available—fail to address any of the relevant questions to sustain the charge of assault, riot or murder. They do not help to answer the question of identity or the question of what the worker on trial did.

8 Cold-hit DNA matches

Another set of cases will help to sharpen the distinction between statistics that are problematic and statistics that are not: cold-hit DNA cases. The peculiarity of these cases is that the evidence consists exclusively of a DNA match. Suppose investigators cannot identify a suspect by means of traditional methods. But they can collet a specimen from the crime scene and run it against a database of genetic profiles. This database can be extensive, containing the genetic profiles of millions of people. If the search yields a match, the person associated with the matching profile becomes a suspect. Since the investigators did not identify the suspect via any means other than the database search, the only incriminating evidence is the DNA match itself. This scenario raises several questions, in particular, can the cold-hit genetic match alone be enough to convict in absence of other evidence?

A lot has been written on cold-hit cases.[25] The core claim of this paper can help to eliminate some of the confusion. First of all, it is worth noting that DNA matches—even in cold-hit cases—are never the only evidence in a case. There is always other evidence. Crime traces always exist. Now suppose the crime traces told us little about what happened. Say the traces could have been left innocently at the scene. In this case, the DNA match would prove nothing. There would be no case against the person who matched the traces. On the other hand, if the cold-hit DNA match—together with the crime traces—can offer a detailed description about what happed, convicting on the basis of the cold-hit DNA match should not be problematic. Part of the confusion here is that if the match, together with the crime traces, can provide the needed information, it would not be a cold-hit match since it would be—strictly speaking—supplemented by other evidence.

But a worry may linger here. Although they are highly discriminating across individuals, genetic profiles are not unique to individuals.[26] Two people—by sheer coincidence—could have the same genetic profile, the same shoe print, etc. Shouldn't that sheer coincidence give us pause before convicting? It certainly should. However, any form of identification evidence leaves open the possibility of a coincidence. Fingerprint evidence is no better.[27] Any identifying trait could be shared by more than one person.

Two answers can be pursued here. First, we might say that it all depends on how likely the sheer coincidence is going to be. We could set a probability threshold we are wiling to live with. Second, we might require that the incriminating evidence, considered as a whole, singles out one individual and rules out all other suspects, albeit still fallibly. This latter option is attractive but perhaps overly demanding. It might be possible to rule out all other suspects one by one when the pool is limited. When the pool of suspects is large—say all people in a county—ruling out all suspects except one could be difficult. Be that as it may, the core claim stands: if the evidence

[25] Among forensic scientists and legal scholars, see NRC (1996), Balding and Donnely (1996), Stein (2005), Allen and Pardo (2007), Roth (2010), Cheng and Nunn (2016). Among philosophers, see De Macedo (2008), Enoch et al. (2012), Pritchard (2015), Smith (2018), Mayo (2018).

[26] On the question of uniqueness of genetic profiles, see Kaye (2013).

[27] For a comparison between fingerprint evidence and genetic evidence, see Zabell (2005).

is sufficiently specific—and cold-hit DNA matches *plus* crime traces can be very specific—it should be able to sustain a criminal accusation.

9 Probability, specificity, availability

It is time to conclude and hints at further lines of inquiry. Much of the argument of this paper focused on the notion of specificity, and the circumstances in which statistical evidence—or other forms of evidence—may fail the specificity test. Specificity is a property of the accusatory narrative, not directly of the evidence itself, but the evidence can be more or less specific to the extent that it can support a more or less specific narrative. I argued that focusing on the specificity of the accusatory narrative helps to justify why a conviction would be unacceptable in cases featuring naked statistical evidence, such as the worker scenario or the department store scenario. These are examples in which statistical evidence fails the specificity test. I also argued that statistical evidence is not objectionable per se. It is only objectionable when there is a mismatch between the degree of specificity of the evidence—more precisely, the degree of specificity of the accusation which the evidence supports—and the degree of specificity that is expected in the accusation. The nurse and price manipulation examples make this point clear. In these cases, there is no mismatch. Although the statistics are not about a specific action located in space and time, the accusation is not intended to be so specific either. At the same time, I showed that statistical evidence—especially in statistics-based identification evidence such as DNA matches—can contribute to support detailed accusatory narratives about what happened.

As I mentioned earlier in the paper, other strategies exist for tackling the puzzle of naked statistical evidence. They focus on (i) the paucity of the evidence; (ii) the statistical natural of the evidence; and (iii) the impossibility of interpreting the standard of proof in criminal cases as a simple probability threshold. My argument is not incompatible with these approaches. Let me consider each one in turn. (i) If the evidence is particularly thin, the accusatory narrative the evidence can support will tend to be rather unspecific. So there is a strong connection between the paucity of the evidence and the limited specificity of the accusation. (ii) Whenever statistical evidence concerns patterns of behavior over time, it will support a narrative that has limited specificity. But, as I remarked at length, the degree of specificity of an accusatory narrative is not set in stone and varies contextually depending on what questions are relevant and the type of crime being prosecuted. So, while the statistical nature of the evidence justifies to some extent our resistance to a conviction in cases of naked statistical evidence, this is not the entire story. The deeper reason for resisting a conviction has to do with the specificity of evidence, not its statistical nature. (iii) Finally, my argument remains neutral on the question whether the standard of proof in criminal cases should be interpreted as a probability threshold. My argument underscores, however, that we should not understand guilt as an abstract proposition. Prosecutors do not aim to establish the defendant's guilt in and of itself. Rather, they aim to establish a certain accusatory narrative from which guilt follows according to the law governing a case. In probabilistic terms, prosecutor should aim to establish that their accusatory

narrative is highly probable on the evidence, not simply that guilt is highly probable on the evidence.

I conclude with two programmatic hypotheses. First, the litigation process at trial can be seen as an attempt to optimize the incidence of these three variables:

⊙ the probability that the accusation (story, theory, narrative) is true;
⊙ the specificity of the accusation; and
⊙ the availability of evidence to establish the accusation.

The optimal equilibrium point requires that criminal accusations should be established with (a) reasonably high probability, (b) reasonable degree of specificity while also (c) taking into account all evidence reasonably available. Any departure from this optimal equilibrium of probability, specificity and availability of the evidence would give rise to a reasonable doubt. What counts as 'reasonable' is hard to make precise, though. Perhaps, the decision about the optimal level should be left to the individual litigants or alternatively laws and regulations should be clear about what is required.

Here is the second programmatic hypothesis. The optimal equilibrium point of the three variables may be different in civil and criminal trials. There is no doubt that accusations should be established with lower probability in civil than criminal trials. But it might also be that accusations in civil trials should be less specific than in criminal trials. It might also be that the search for evidence should be less demanding in civil than criminal trials. If so, it follows that statistical evidence—even when it is not very specific and merely establishes a tendency or a risk—should be considered more favorably in civil trials.

Both these hypotheses are tentative. Their examination is left for another time.

References

Allen, R. J. (2010). No Plausible Alternative to a Plausible Story of Guilt as the Rule of Decision in Criminal Cases. In J. Cruz & L. Laudan (Eds.), *Prueba y Esandares de Prueba en el Derecho*. Instituto de Investigaciones Filosoficas-UNAM.

Allen, R. J. (2021). Naturalized epistemology and the law of evidence revisited. *Quaestio Facti*, *2*, 253–284.

Allen, R. J., & Leiter, B. (2001). Naturalized epistemology and the law of evidence. *Virginia Law Review*, *87*, 1491–1550.

Allen, R. J., & Pardo, M. S. (2007). The problematic value of mathematical models of evidence. *Journal of Legal Studies*, *36*(1), 107–140.

Allen, R. J., & Pardo, M. S. (2019). Relative plausibility and its critics. *International Journal of Evidence and Proof*, *23*(1/2), 5–59.

Balding, D. J., & Donnely, P. (1996). Evaluating DNA Profile evidence when the suspect is identified through a database search. *Journal of Forensic Science*, *41*(4), 603–607.

Blome-Tillmann, M. (2015). Sensitivity, causality, and statistical evidence in courts of law. Thought A. *Journal of Philosophy*, *4*(2), 102–112.

Bolinger, R. J. (2020). The rational impermissibility of accepting (some) racial generalizations. *Synthese*, *197*, 2415–2431.

Buchak, L. (2014). Belief, credence, and norms. *Philosophical Studies*, *169*(2), 285–311.

Burns, R. P. (2004). The distinctiviness of trial narratives. In A. Duff, L. Farmer, S. Marshall, & V. Tadros (Eds.), *The trial on trial (VOL 1): truth and due process*. Hart Publishing.

Carnap, R., & Bar-Hillel, Y. (1952). An outline of a theory of semantic information. *Technical Report MIT*.

Cheng, E. K. (2013). Reconceptualizing the burden of proof. *Yale Law Journal*, *122*(5), 1254–1279.

Cheng, E. K., & Nunn, G. A. (2016). DNA, blue bus, and phase changes. *The International Journal of Evidence and Proof*, *20*(2), 112–120.

Cohen, J. L. (1977). *The Probable and the Provable*. Oxford University Press.

Colyvan, M., Regan, H. M., & Ferson, S. (2001). Is it a crime to belong to a reference class? *Journal of Political Philosophy*, *9*(2), 168–181.
Dahlman, C. (2020). Naked statistical evidence and incentives for lawful conduct. *International Journal of Evidence and Proof*, *24*(2), 162–179.
De Macedo, C. (2008). Guilt by statistical association: Revisiting the prosecutor's fallacy and the interrogator's fallacy. *Journal of Philosophy*, *105*(5), 320–332.
Di Bello, M. (2013). *Statistics and Probability in Criminal Trials: The Good, the Bad and the Ugly.* Ph.D. thesis, Stanford Univerity.
Di Bello, M. (2019). Trial by statistics: Is a high probability of guilt enough to convict? *Mind*, *128*(512), 1045–1084.
Di Bello, M., & O'Neil, C. (2020). Profile evidence, fairness and the risk of mistaken convictions. *Ethics,130*(2), 147–178.
Duff, A. (2001). *Punishment, Communication and Community*. Oxford University Press.
Ebert, P. A., Smith, M., & Durbach, I. (2018). Lottery judgments: A philosophical and experimental study. *Philosophical Psychology*, *31*(1), 110–138.
Enoch, D., Fisher, T., & Spectre, L. (2012). Statistical evidence, sensitivity, and the legal value of knowledge. *Philosophy and Public Affairs*, *40*(3), 197–224.
Feinberg, S. E., & Kaye, D. H. (1991). Legal and statistical aspects of some mysterious clusters. *Journal of the Royal Statistical Society Series A (Statistics in Society)*, *154*, 61–74.
Fitelson, B. (2006). Likelihoodism, Bayesianism, and relational confirmation. *Synthese*, *16*(47), 1–22.
Floridi, L. (2004). Outline of a theory of strongly semantic information. *Minds and Machines*, *14*, 197–222.
Gardiner, G. (2018). Legal burdens of proof and statistical evidence. In D. Coady & J. Chase (Eds.), *Routledge Handbook of Applied Epistemology*. Routledge.
Gardiner, G. (2019). The reasonable and the relevant: Legal standards of proof. *Philosophy and Public Affairs*, *47*(3), 288–318.
Gardiner, G. (2020). Profiling and proof: Are statistics safe? *Philosophy*, *95*(2), 161–183.
Groenendijk, J., & Stokhof, M. (1997). Questions. In J. van Benthem (Ed.), *Handbook of Logic and Language*. Elsevier and MIT Pess.
Haack, S. (2014). *Evidence Matters: Science, Proof, and Truth in the Law*. Cambridge University Press.
Hedden, B., & Colyvan, M. (2019). Legal probabilism: A qualified defence. *Journal of Political Philosophy*, *27*(4), 448–468.
Ho, H. L. (2008). *Philosophy of Evidence Law*. Oxford University Press.
Kaye, D. H. (1979). The paradox of the gatecrasher and other stories. *The Arizona State Law Journal, 1979*(1), 101–110.
Kaye, D. H. (1986). Do we need a calculus of weight to understand proof beyond a reasonable doubt? *Boston University Law Review*, *66*, 657–672.
Kaye, D. H. (2013). Beyond uniqueness: the birthday paradox, source attribution and individualization in forensic science. *Law, Probability and Risk*, *12*(1), 3–11.
Koehler, J. J. (2002). When do courts think base rate statistics are relevant? *Jurimetrics Journal*, *42*, 373–402.
Krauss, S. F. (2020). Against the alleged insufficiency of statistical evidence. *Florida State University Law Review*, *47*, 801–825.
Laudan, L. (2011). The rules of trial, political morality and the costs of error: Or, Is proof beyond a reasonable doubt doing more harm than good? In G. Leslie & L. Brian (Eds.), *Oxford Studies in Philosophy of Law* (Vol. 1). Oxford University Press.
Littlejohn, C. (2020). Truth, knowledge, and the standard of proof in criminal law. *Synthese*, *197*, 5253–5286.
Loftus, E. F. (1996). *Eyewitness Testimony (revised edition)*. Harvard University Press.
Malcom, B. G. (2008). Convictions predicated on DNA evidence alone: How reliable evidence became infallible. *Columbia Law Review*, *38*(2), 313–338.
Mayo, D. (2018). *Statistical Inference as Severe Testing*. Cambridge University Press.
Meester, R., Collins, M., Gill, R., & van Lambalgen, M. (2006). On the (ab)use of statistics in the legal case against the nurse Lucia de B. *Law, Probability and Risk*, *5*(3–4), 233–250.
Moss, S. (2021). Knowledge and legal proof. In T. Szabo Gendler & J. Hawthorne (Eds.), *Oxford studies in epistemology* (Vol. 7). Oxford University Press.
Nance, D. A. (2016). *The burdens of Proof: Discriminatory Power, Weight of Evidence, and Tenacity of Belief*. Cambridge University Press.
Nelkin, D. N. (forthcoming). Rational belief and statistical evidence: Blame, bias, and the law. In D. Igor (Ed.), *The Lottery Paradox*. Cambridge University Press.

Nesson, C. R. (1979). Reasonable doubt and permissive inferences: The value of complexity. *Harvard Law Review*, *92*(6), 1187–1225.
Niedermeier, K. E., Kerr, N. L., & Messeé, L. A. (1999). Jurors' use of naked statistical evidence: Exploring bases and implications of the Wells effect. *Journal of Personality and Social Psychology,76*(4), 533–542.
NRC. (1996). *The Evaluation of Forensic DNA evidence*. National Academy Press.
Nunn, A. G. (2015). The incompatibility of due process and naked statistical evidence. *Vanderbilt Law Review*, *68*(5), 1407–1433.
Papineau, D. (2021). The disvalue of knowledge. *Synthese*, *198*, 5311–5332.
Pennington, N., & Hastie, R. (1991). A cognitive theory of juror decision making: the story model. *Cardozo Law Review*, *13*, 519–557.
Picinali, F. (2016). Base-rates of negative traits: Instructions for use in criminal trials. *Journal of Applied Philosophy*, *33*(1), 69–87.
Porat, A., & Posner, E. (2012). Aggregation and law. *Yale Law Journal*, *122*, 2–69.
Pritchard, D. (2015). Risk. *Metaphilosophy*, *46*(3), 436–461.
Pundik, A. (2017). Freedom and generalisation. *Oxford Journal of Legal Studies*, *37*(1), 189–216.
Redmayne, M. (2015). *Character in the Criminal Trial*. Oxford University Press.
Ross, L. (2021). Rehabilitating statistical evidence. *Philosophy and Phenomenological Research*, *102*(1), 3–23.
Roth, A. (2010). Safety in numbers? Deciding when DNA alone is enough to convict. *New York University Law Review*, *85*(4), 1130–1185.
Schauer, F. (2003). *Profiles, Probabilities, and Stereotypes*. Belknap Press.
Schauer, F. (2021). *Statistical evidence and the problem of specification*. Unpublised Manuscript.
Schmalbeck, R. (1986). The trouble with statistical evidence. *Law and Contemporary Problems*, *49*(3), 221–236.
Schoeman, F. (1987). Statistical vs. direct evidence. *Noûs*, *21*(2), 179–198.
Smith, M. (2018). When does evidence suffice for conviction? *Mind*, *127*(508), 1193–1218.
Staffel, J. (2016). Beliefs, buses and lotteries: Why rational belief can't be stably high credence. *Philosophical Studies,173*, 1721–1734.
Stein, A. (2005). *Foundations of Evidence Law*. Oxford University Press.
Thomson, J. J. (1986). Liability and individualized evidence. *Law and Contemporary Problems*, *49*(3), 199–219.
Tribe, L. H. (1971). A further critique of mathematical proof. *Harvard Law Review*, *84*(8), 1810–1820.
Urbaniak, R. (2018). Narration in judiciary fact-finding: a probabilistic explication. *Artificial Intelligence and Law*, *26*(4), 345–376.
van Koppen, P., & Mackor, A. R. (2020). A scenario approach to the Simonshaven case. *Topics in Cognitive Science*, *12*(4), 1132–1151.
Wasserman, D. T. (1991). The morality of statistical proof and the risk of mistaken liability. *Cardozo Law Review*, *13*, 935–976.
Weinstein, J. B., & Dewsbury, I. (2006). Comment on the meaning of 'proof beyond a reasonable doubt'. *Law, Probability and Risk,5*(2), 167–173.
Wells, G. L. (1992). Naked statistical evidence of liability: Is subjective probability enough? *Journal of Personality and Social Psychology*, *62*(5), 752–793.
Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest*, *18*(1), 10–65.
Zabell, S. L. (2005). Fingerprint evidence. *Journal of Law and Policy*, *13*(1), 143–179.