Synthese (2019) 196:4253–4278
https://doi.org/10.1007/s11229-017-1654-y

Meno's paradox and medicine

Nicholas Binney[1]

Received: 31 May 2017 / Accepted: 8 December 2017 / Published online: 26 December 2017

Abstract The measurement of diagnostic accuracy is an important aspect of the evaluation of diagnostic tests. Sometimes, medical researchers try to discover the set of observations that are most accurate of all by directly inspecting diseased and not-diseased patients. This method is perhaps intuitively appealing, as it seems a straightforward empirical way of discovering how to identify diseased patients, which amounts to trying to correlate the results of diagnostic tests with disease status. I present three examples of researchers who try to produce definitive diagnostic criteria by directly inspecting diseased and not diseased patients. Despite this method's intuitive appeal, I will argue that it is impossible to carry out. Before researchers can inspect these patients to discover definitive diagnostic criteria, they must be able to distinguish diseased and not-diseased patients; and they do not know how to do this, because this is what they are trying to discover. I suspect the intuitive appeal of directly inspecting patients makes this difficult to appreciate. To counter this difficulty, I present this problem as a manifestation of 'Meno's paradox', which was described in classical antiquity, and of 'the problem of nomic measurement', described more recently. Considering these philosophical problems may help researchers address the methodological issues they face when evaluating diagnostic tests.

Keywords Diagnostic accuracy · Meno's paradox · The problem of nomic measurement · Medical epistemology

✉ Nicholas Binney
nb357@exeter.ac.uk

[1] EGENIS – Centre for the Study of Life Sciences, University of Exeter, Byrne House, St. German's Road, Exeter, Devon EX4 4PJ, UK

1 Introduction

The question of how best to detect disease is central to medical practice. This question is most often addressed by measuring the *accuracy* of different methods of detecting disease (Kennedy 2016). When evaluating diagnostic tests in this way, doctors are concerned with determining the true disease status of a patient, and want to know how well a test discriminates between diseased and not-diseased patients. The diagnostic accuracy of a test (or battery of tests and observations) is "the ability of the test to discriminate between patients with and without the target condition" (Reitsma et al. 2009, p. 797).

The accuracy of a method of detecting disease is measured by comparing the performance of a diagnostic practice under evaluation, which is commonly referred to as the "index test", to that of a diagnostic practice which is trusted to deliver an accurate result, which is commonly referred to as the "gold standard" or the "reference standard" (Knottnerus et al. 2009; Newman and Kohn 2009, p. 99). An index test that returns results that are in close agreement with the reference standard is considered accurate and therefore good. Should the index test return positive results in all patients with the disease it is said to be a perfectly "sensitive" test. Should the index test return negative results in all patients without the disease it is said to be a perfectly "specific" test. There is a vast medical literature that measures the accuracy of different methods of detecting disease in this way, and that presents arguments about how good certain diagnostic practices are at detecting particular diseases based on these results (Knottnerus and Buntinx 2009; Newman and Kohn 2009).

Many scholarly works have addressed the question of how to evaluate diagnostic practices (Mackenzie 1916; Kahn 1942; Feinstein 1967; Wulff 1976; Gøtzsche 2007). Despite this, medical researchers still argue that the "Poor quality of diagnostic studies is a recognised problem" (Fontela et al. 2009), and that "the theory and methodology of diagnostic research still lags substantially behind that of research into the effectiveness of treatment" (Knottnerus and Buntinx 2009, p. 11). As identified by Kennedy (2016; see also Mebius et al. 2016), philosophers have also paid almost no attention to how diagnostic practices are evaluated in medical practice. One aim of this paper is to draw attention to the need for researchers, clinicians and philosophers to pay more attention to the methodology of diagnostic research. I will argue that this is the case because many arguments that evaluate the accuracy of diagnostic practices are flawed.

Due to concerns about the quality of studies designed to measure the accuracy of diagnostic tests, several sets of guidance have been prepared to advise researchers on how to conduct and report their research. These sets of guidance include editions of the STARD (standards for reporting diagnostic accuracy) and the QUADAS (quality assessment for studies of diagnostic accuracy) (Whiting et al. 2011, 2003; Bossuyt et al. 2003, 2015). These sets of guidance rely on there being agreed upon reference standards against which to assess the performance of index tests. They do not, however, offer guidance about what to do when there is no consensus about what the reference standard for a disease is. When doctors do not agree about which patients have a disease, how does one assess different reference standards so that the best one may be chosen and used?

A recent paper, entitled "A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard", has systematically reviewed proposed methods for evaluating diagnostic tests in the absence of a gold standard (Reitsma et al. 2009). These potential solutions include a number of interesting techniques, which deserve philosophical attention. These techniques include using multiple tests to try to more accurately diagnose disease, trying to adjust results to account for the imperfections of the reference standard, appealing to a panel of experts to construct a reference standard, and latent class analysis. However, in this review one important method of measuring diagnostic accuracy is omitted—*the direct inspection of diseased and not-diseased patients to identify distinguishing characteristics*. This amounts to trying to make correlations between observations and disease status in order to determine what the gold standard should be. I say that this is an important method because many doctors and researchers try to do this. Despite these attempts, I will argue that this is impossible.[1] I suspect Reitsma et al. (2009) recognise this, as they say that the normal techniques for evaluating diagnostic tests do not apply to situations where the status of a test as the most accurate test of all is questioned. "One special situation is where the index test is proclaimed to be better than the reference standard" (Reitsma et al. 2009, p. 803). These researchers do not, however, explain why this situation is special. Perhaps they do not do this because they believe that it is obvious that an index test cannot be found to be more accurate than the reference standard, as the reference standard is the standard by which the accuracy of the index test is judged. An arrow cannot be closer to the centre of a target than the centre of the target. Even so, that the evaluation of a reference standard is a special epistemic situation has not been appreciated by many medics, who continue to try to identify the most accurate way of diagnosing disease by the direct inspection of diseased and not-diseased patients. This paper will discuss three examples of researchers who endorse this approach. Together these examples not only support the thesis that this approach to the evaluation of diagnostic practices is widespread, but also provide a sense of the variety of research contexts to which this problem applies. I draw attention to a profound problem with this methodology, which I describe as a manifestation of a very ancient philosophical problem, first articulated in Plato's *Meno*.

My first example, discussed in Sect. 2, is drawn from the medical literature on the diagnosis of eosinophilia–myalgia syndrome (EMS) (Hertzman et al. 2001). These researchers inspected a group of patients with the syndrome to see whether or not a set of diagnostic criteria they suggested was able to correctly distinguish diseased patients from not-diseased patients. They found these criteria to be accurate, and recommended them for use. I believe these researchers were primarily interested in producing a set of diagnostic criteria that could be used by non-experts to select the same patients as EMS experts, and that their methods are suited to this limited goal taken on its own.

[1] Of course, if everyone is in agreement about what the which patients have a disease, then some distinguishing characteristics can be identified by direct inspection. The problem I identify here only arises when there is disagreement about which patients have a disease. By distinguishing characteristics, I mean the set of all observations that are used to distinguish diseased and not-diseased patients. For example, these may be symptoms reported by the patient, signs observed by the doctor at physical examination, blood test results, histopathology, results of post mortem examinations, results of diagnostic imaging, and the results of therapeutic trials.

However, these researchers also claim to have measured the diagnostic accuracy of these criteria as objectively as possible, and they recommended that their methods should be used by other researchers evaluating diagnostic practices in the absence of a gold standard; and these claims are problematic.

The difficulties with this method are more apparent from my second example, discussed in Sect. 3, which is drawn from the medical literature on fibromyalgia (Wolfe et al. 1990). These researchers attempt to establish a set of diagnostic criteria as the most accurate available by direct inspection, using similar methods to those recommended by Hertzman et al. (2001). These researchers are, in effect, trying to discover the criteria that can be used to define the groups of patients with and without a certain target condition by correlating diagnostic tests results and disease status. Although this may seem like a sensible empirical approach, it is an impossible task. The trouble with this approach is that in order to inspect the group of patients with a disease to see what distinguishes them from not-diseased patients, researchers must already know how to distinguish diseased patients from not-diseased patients. If researchers do not know how to do this at the beginning of the study, they will be unable to identify diseased and not-diseased patients, which they need to do to carry out their research. Even though this may seem like an obvious problem that should not manifest in practice, I show here that researchers do fall into this trap.[2]

In Sect. 4, I argue that this problem is a manifestation of an ancient philosophical problem discussed in Plato's *Meno*, referred to as *Meno's paradox* (Ebrey 2014). The precise meaning of the passages where Plato sets out this problem is debated by scholars, and I will not engage with these debates here. It is not my intention to contribute to this scholarship by re-reading the *Meno* in the light of the epistemic problems encountered in medical practice. Even though Meno's paradox may be ambiguous, in my view Plato put his finger on an important epistemic problem, and I believe that considering this part of the *Meno* helps to clarify problems encountered in medical practice. I also link Meno's paradox with another similar and perhaps less ambiguous formulation of the problem, identified by Chang (2004) when discussing the measurement of temperature—*the problem of nomic measurement.* Attending to these philosophical problems throws the difficulties encountered in the medical literature into sharp relief, and understanding them will help doctors identify and articulate these difficulties when they arise in other areas of medicine.

In Chang's discussion of temperature, the problem of nomic measurement manifests when researchers are trying to discover how to measure a metaphysical entity, which it is not possible to observe directly (2004, p. 59). Some medics do seem to think of diseases in this way. Claims that no diagnostic observation is perfectly accurate occur in medical literature (Cardoso et al. 2014, p. 29; Okeh and Okoro 2012; Duggan 1992; Versi 1992). Many prefer the term "reference standard" to the term "gold standard" precisely because it reflects the view that tests are imperfect (Knottnerus and Muris 2009, p. 50; Weinstein et al. 2005, p. 18). If all possible observations are always imperfect indicators of a disease, then that disease is not itself observable.

[2] Two of the three examples I explore here are syndromes. It may be that the paradoxical problems I am highlighting here are particularly common in discussions about how syndromes are diagnosed. Further empirical work is needed to answer this question.

This view of disease fits with Chang's account of the problem of nomic measurement, but other views of disease are also available. Some argue that diseases are reductions of biological function (Ereshefsky 2009), so the identification of reduced biological function would count as the direct observation of a disease. There is also disagreement about whether diseases are objective, real and natural kinds, or if they are subjective, constructed and artificial kinds (Ereshefsky 2009; Simon 2017). Regardless of whether diseases are taken to be observable or unobservable, objective or subjective, the problem I identify here may still appear, just so long as researchers do not know (or are arguing about) how to recognise patients with the disease.

It is particularly important to appreciate these problems when there is a dispute about what the characteristics by which patients with a disease can be identified are. My third example, discussed in Sect. 5, is drawn from the medical literature on the diagnosis of rickets in infants (Slovis and Chapman 2008b). I focus on a dispute about how patients with rickets can be identified, and specifically on whether the classic radiographic signs of rickets need to be present for a diagnosis to be made. The arguments deployed in this dispute are less clearly structured than those used in my other examples. Nevertheless, I will show that Meno's paradox still manifests in this literature, and argue that this discussion would be improved by recognition of this problem.

Some doctors suggest that when there is no agreed upon gold standard, "the diagnostic accuracy paradigm may be abandoned in favour of alternative methods for evaluating diagnostic tests" (Reitsma et al. 2009, p. 803). In Sect. 6, I close with a brief discussion of some alternative strategies for evaluating diagnostic tests, as I do not want to signal that reference standards cannot be evaluated.

2 Eosinophilia–myalgia syndrome

Eosinophilia–myalgia syndrome (EMS) is a disease marked by dramatic increases of a certain form of white blood cell, the eosinophil, in the patient's body, accompanied by severe muscle pain (myalgia), fever, skin changes and respiratory symptoms (Bulpitt et al. 1990, p. 918). It was identified following an outbreak of the syndrome that was associated with the consumption of a certain brand of nutritional supplement containing tryptophan in the late 1980s. Hertzman et al. (2001, p. 2302) report that EMS has proved difficult to diagnose in epidemiological studies of the disease, and that suggested diagnostic criteria had not been validated. They sought to produce and validate a set of diagnostic criteria by measuring their diagnostic accuracy.

These researchers first produced a set of criteria that they believed could serve as the most accurate diagnostic criteria for EMS. They did this by compiling a list of 45 clinical observations and test results "considered important in the diagnosis of EMS and related disorders, and then, by consensus, reduced this list to 10 elements". So these 10 elements were considered the most important for the diagnosis of EMS. These 10 elements were assembled into two "patterns" considered characteristic of EMS.[3]

[3] Pattern 1 is the presence of eosinophilia, myalgia and at least one of rash, edema, pulmonary involvement, or neuropathy, occurring within a six-month period. Pattern 2 is the presence of fasciitis, myopathy, and

These patterns together functioned as criteria for the diagnosis of EMS, as the authors recommended that "EMS can be diagnosed if either pattern 1 or 2 is satisfied", so long as other potential causes of illness were first ruled out. "EMS should not be diagnosed in the presence of trichinosis, vasculitis, or any other documented infectious, allergic, neoplastic, connective tissue or other type of disease that could adequately explain the clinical manifestations" (Hertzman et al. 2001, p. 2303).

So Hertzman et al. (2001) produced a set of criteria for the diagnosis of EMS that they believed should be accurate. But how did they check that this was actually the case? They needed a set of cases with known disease status to which they could apply their criteria, to see if their criteria correctly classified these patients. As no formal and accepted criteria existed to select cases who definitely did and did not have EMS, Hertzman et al. (2001) invited a panel of experts to select cases to serve as this set of cases with known disease status. "Because EMS lacks discrete pathognomonic features, we thought that the best surrogate gold standard for testing the criteria would be a set of reports of EMS cases and noncases that were generated and validated by an external panel of experts" (Hertzman et al. 2001, p. 2302). Each expert was asked to provide five cases that they diagnosed with EMS, four cases without EMS but diagnosed with conditions resembling EMS, and one possible but uncertain case of EMS (Hertzman et al. 2001, p. 2303). These cases were reviewed by the other panel members, and if 75% of the panel agreed about this patient's diagnosis then the case was retained in the set with the agreed diagnosis. This procedure produced a population of cases, 50 of whom were taken to have the disease, and 35 of whom were taken not to have EMS, but instead to have a condition resembling EMS. This set of cases with known disease status was referred to as "the gold standard set" (Hertzman et al. 2001, p. 2303). Although these researchers thought that their reliance on the opinions of a panel of experts is in some sense arbitrary, they still claimed that their gold standard set was as objective as possible. "Although any gold standard, regardless of method of construction, would be arbitrary, we believed that this approach would result in the most appropriate objective standard possible" (Hertzman et al. 2001, p. 2302).

Hertzman et al. (2001) then applied their diagnostic criteria to this gold standard set. They found that each researcher was in close agreement with the others with regard to which patients in the gold standard set satisfied their criteria. They found that their criteria returned positive results in 44 of the 50 cases of EMS in the gold standard set, showing a sensitivity of 88%. Their criteria returned negative results in 34 of the 35 cases of not EMS in the gold standard set, showing a specificity of 97%. As 78 out of 85 results were correct, the overall accuracy of their criteria was measured as 92%. Hertzman et al. (2001) conclude as follows:

> The proposed criteria are accurate and reproducible, and can be used in future clinical investigations of the eosinophilia-myalgia syndrome. The new strategy and methods developed for this challenge can be valuable for solving analogous

Footnote 3 continued
myalgia or muscle cramps; or alternatively any three of fasciitis, myopathy, neuropathy, or eosinophilia (Hertzman et al. 2001, p. 2303).

> problems in constructing criteria for other clinical disorders (Hertzman et al. 2001, p. 2301).

Hertzman et al.'s (2001) approach to producing these diagnostic criteria is carefully considered. They put in place a number of formal methods to try to ensure that the test of their criteria is fair. In particular, they are at pains to emphasize that the expert panel who produced the gold standard set was independent of the researchers who applied the candidate diagnostic criteria. They also emphasize that the application of these criteria was made without knowledge of the gold standard diagnosis for each case. This was done so that the diagnostic decisions of the expert panel should not inform the diagnostic decisions of the researchers. "The gold standard case sets were defined and the criteria were interpreted independently, so the results of one process did not influence the other"[4] (Hertzman et al. 2001, p. 2306).

Despite these researchers' efforts to keep the construction of the gold standard set and the candidate diagnostic criteria independent, these processes are only independent in a very limited sense. In standard studies of diagnostic accuracy, researchers are not trying to measure the accuracy of the definitive diagnostic criteria for a disease, which are the criteria used to define the group of patients with the disease (the gold or reference standard). Rather, researchers are trying to measure the accuracy of some supplementary test that is not used to define the group of patients with the disease (the index test). Researchers typically want to know how accurate these supplementary tests are because the definitive diagnostic criteria cannot be applied for some reason. Perhaps these criteria too expensive to apply, or too invasive, or involve knowledge of some future event such as the development of particular symptoms, or are post mortem observations. In this case, however, researchers are trying to produce definitive diagnostic criteria for use in clinical and research practice. Recall that the "elements" from which Hertzman et al. (2001) fashioned the two patterns of EMS were deemed to be the ten *most important* elements for the diagnosis of EMS. These elements that make up the candidate diagnostic criteria are not supplementary observations. The candidate diagnostic criteria are not independent of the observations used to produce the gold standard set of patients— *they are the observations used to produce the gold standard set of patients*.

Hertzman et al. (2001) do not describe which elements were judged to be important by the expert panel as they assembled the gold standard set of cases. Even if the elements used by the expert panel were the same as those judged to be the ten most

[4] Both STARD and QUADAS require researchers to check that the gold standard and index test are independent of each other. Independence in this context has at least two meanings. The first is that the researchers applying the index test should be blinded to the results of the reference standard, and vice versa, so that this knowledge does not influence their assessment of the test result. The second is that the index test should not form a component part of the process by which the disease status of the patient is determined, because this will mean that a patient that satisfies the gold standard will automatically be more likely to test positive with the index test. The incorporation of the index test into the gold standard may lead to the appearance that the index test is more accurate that it actually is, a situation referred to as "incorporation bias" (Worster and Carpenter 2008; Newman and Kohn 2009, p. 99). The solution to this epistemic problem is to separate the gold standard and the index test, which is simple to do under normal circumstances when an index test is evaluated against a gold standard. However, when it is the gold standard that is under evaluation, the epistemic challenge is rather more substantial.

important by Hertzman et al. (2001), there is no way of telling if they were organised into the same two patterns. Even so, in the absence of evidence to the contrary, it seems sensible to assume that the patterns of elements used by the expert panel were very similar to those proposed by Hertzman et al. (2001). This is because Hertzman et al. (2001) judged these panellists to be fellow experts on EMS, and thus most likely agreed with their views on EMS.

Readers of Hertzman et al. (2001) might be forgiven for believing that the accuracy of the diagnostic criteria proposed by these researchers had been established by correlating the results of these observations with the disease status of patients. The accuracy of these candidate diagnostic criteria is not established by empirical observation. Rather, they have (at least tacitly) been assumed by the expert panel to carry out this study. The association of these patterns of elements and EMS is not established by empirical observation in this study. Rather this association is a precondition for the study to be carried out at all.

I do not want to be too critical of Hertzman et al. (2001). In the particular context of the diagnosis of EMS, these researchers achieve most of their goals. Hertzman et al. (2001) do not suggest that experts are not able to recognise EMS when they see it. The diagnostic criteria they propose are probably not designed to assist experts in making the diagnosis. These criteria appear designed to assist doctors who are not experts in EMS in their clinical work and whilst doing epidemiological research into the disease. Hertzman et al. (2001) sought to develop a set of diagnostic criteria that, when applied by non-experts, could mimic the diagnostic performance of an expert.[5] It may be that Hertzman et al. (2001) sought to replicate the results of the very complex and perhaps tacit decision-making process used by experts in EMS in a simple and explicit set of diagnostic criteria. This research does contribute to these goals. Even so, Hertzman et al. (2001) do not limit themselves to claiming that their diagnostic criteria faithfully reproduce expert performance. They claim that their diagnostic criteria are accurate at detecting EMS, that their gold standard set of patients was as objective as possible, and that their methods should be used to assess the accuracy of diagnostic criteria in other situations where there is no accepted gold standard.

3 Fibromyalgia

The problematic nature of this methodology for evaluating diagnostic criteria is apparent from studies that do something like what Hertzman et al. (2001) suggest, but try to use the results of their study to inform expert opinion about how to diagnose disease. Consider the 1990 American College of Rheumatology guidelines on the diagnosis of fibromyalgia (Wolfe et al. 1990). Fibromyalgia is a condition characterized by widespread and unexplained pain, particularly in response to pressure applied to certain areas of the body described as “tender points”. The diagnosis of fibromyalgia is

[5] Hertzman et al. (2001, p. 2305) also emphasise that they expect their suggested criteria to change as more is learned about EMS. They do not claim to be certain that their diagnostic criteria are as accurate as they say they are.

contested, with some doctors denying that it is a discrete disease entity at all (Wessely and Hotopf 1999; Cohen and Quintner 1993).

In contrast to Hertzman et al. (2001) in the case of EMS, these researchers emphasized the diversity of different diagnostic criteria that were employed by experts with a special interest in fibromyalgia (Wolfe et al. 1990, p. 161). Two of the aims of this research are stated as "to provide a consensus definition of fibromyalgia" and "to establish new criteria for the classification of fibromyalgia" (Wolfe et al. 1990, p. 161). Wolfe et al. (1990) are clear that their aim is to contribute to a discussion amongst experts about how to identify patients with fibromyalgia.

Wolfe et al.'s (1990) methodology is similar to that suggested by Hertzman et al. (2001). They invited a group of experts to put forward a set of 293 patients with and 265 without the disease.[6] They then trained a group of "independent assessors" to interview and examine these patients. These assessors collected information deemed relevant to a diagnosis of fibromyalgia—including information about sleeping patterns, morning stiffness, the presence of irritable bowel syndrome, the presence of widespread pain, response to pressure applied at specific sites (the "tender points"), and sensitivity to pain (measure with a dolorimeter) (Wolfe et al. 1990, pp. 161–162). Statistical analyses were made to identify the combinations of elements which could serve as diagnostic criteria.[7] "[V]arious combinations of symptoms were tested in combination with different levels of tender point positivity to identify which items or groups of items performed best" (Wolfe et al. 1990, p. 163). Wolfe et al. (1990) concluded that the presence of widespread pain and tenderness at at least 11 out of 18 tender points were sufficiently accurate to be used as diagnostic criteria for fibromyalgia. "The newly proposed criteria for the classification of fibromyalgia are (1) widespread pain in combination with (2) tenderness at 11 or more of the 18 specific tender point sites" (Wolfe et al. 1990, p. 160). Wolfe et al. (1990) also found that these criteria were more accurate than any other combination of elements that they explored, and more accurate than other candidate diagnostic criteria that had been suggested by other researchers in the past. "Various combinations of tender point levels and groups of symptoms were tested (as in the criteria described by Yunus et al.), but none proved to be as sensitive, specific, and accurate as the combination of widespread pain and 11 of 18 tender points" (Wolfe et al. 1990, p. 170). Wolfe et al. (1990) therefore argued that the diagnostic criteria they advocate should be accepted by experts because they are the most accurate of all.

Wolfe et al. (1990) were concerned about the prospect that their conclusions may simply be the result of the assumptions made to carry out the study, as was discussed above in the case of EMS. These researchers were concerned about this because they recognised that other studies had fallen into this trap, and deployed circular arguments:

[6] Similarly to Hertzman et al. (2001), the patients put forward by Wolfe et al. (1990) as not having fibromyalgia were not clinically well, but rather judged to have another condition with a somewhat similar presentation.

[7] This is a difference between the methodologies of Hertzman et al. (2001), who used their gold standard set of patients to test the performance of candidate diagnostic criteria, and Wolfe et al. (1990), who inspected patients with and without disease to produce their diagnostic criteria. I consider both of these practices to be instances of inspecting groups of diseased and not-diseased patients to identify distinguishing characteristics.

> Even so, there were serious methodologic problems with these criteria sets. Most had not been tested clinically, and none had been tested beyond the centers in which they were designed. No studies had used blinding. Most often, the definitions for the historical features, and even the physical examination features, were imprecise. The most important concern about the criteria, however, was that they tended to be circular; that is, the criteria confirmed the definition of fibromyalgia that was held by the investigators who developed them, a confirmation that might have been assisted by the unblinded status. It was with these objections in mind that the committee undertook the current study (Wolfe et al. 1990, p. 161).

Wolfe et al. (1990) take steps to try to prevent the opinions of individual experts influencing the outcome of their investigations. This is why they invited many different experts to contribute cases to the study, so that the attitudes of one expert would not be over-represented in the study:

> The committee was aware that the way the investigators perceived the syndrome might affect the diagnosis and the sensitivity and specificity of the diagnostic criteria. To reduce diagnosis-criteria circularity, a "consensus" diagnosis of fibromyalgia was obtained by inviting the participation of all centers in Canada and the United States who had a known interest in fibromyalgia (Wolfe et al. 1990, p. 169).

Sadly, this precaution does not resolve the issue at hand. According to Wolfe et al. (1990), different researchers had different opinions about how to diagnose fibromyalgia. Even so, the opinions of different researchers were not vastly different to those of others. Wolfe et al. (1990) place the views of different researchers of a spectrum. At one end are researchers who are happy to make the diagnosis of fibromyalgia in patients who have a high number of tender points and widespread pain, even in the absence of other symptoms. At the other are researchers who are happy to make the diagnosis in patients with as few as two tender points, so long as other symptoms are present in addition to widespread pain. In the middle are researchers who require both additional symptoms and high counts of tender points (Wolfe et al. 1990, p. 161). So, all parties adopted the view that widespread pain, the presence of tender points and other symptoms were important for the diagnosis of fibromyalgia when selecting cases for this study. As has been noted by other commentators on these studies (Wessely and Hotopf 1999, p. 429; Cohen and Quintner 1993), it is therefore not surprising that widespread pain, tender points and other symptoms were found to be important for the diagnosis of fibromyalgia.[8] The diagnostic criteria suggested by Wolfe et al.

[8] Barker's (2005) work on the sociology of fibromyalgia discusses Wolfe et al. (1990), which were the American College of Rheumatology's (ACR) guidelines for the diagnosis of fibromyalgia at one point. Barker reports that many studies of diagnostic accuracy in fibromyalgia deploy circular arguments. "Moreover, the ACR'S analysis was built on a methodological flaw, set in motion by Smythe, and reproduced in every subsequent FMS diagnostic study. FMS is a tautology, tender points both define and substantiate its existence" (2005, p. 25). Barker does not, however, try to explain how such circular arguments get produced. Neither does she identify that Wolfe et al. (1990) were aware of the problem of circular arguments, and took steps to avoid this trap. It is worth exploring why researchers keep falling into this trap, even though they try to avoid it.

(1990) are not the outcome of the careful observation of patients. Rather, it is merely a reflection of the aggregate opinion of a community of doctors who take a special interest in patients with widespread pain. This research does not provide evidence about the characteristics of patients who truly have this target condition, if there are any such patients at all. "While these definitions did improve reliability, the reasoning underlying both papers was essentially circular, and certainly did not provide any evidence for the validity of the concept. Instead, 'thus a pain syndrome is said to define itself"' (Wessely and Hotopf 1999, p. 429; citing Cohen and Quintner 1993).[9]

I offer no opinion about whether EMS and fibromyalgia as described by Hertzman et al. (2001) and Wolfe et al. (1990) are real, nor about whether these categories are valuable or not. I only want to draw attention to the problem that arises when researchers try to discover definitive diagnostic criteria by trying to correlate observations with disease status. Having described this problem in medical practice today, I will now cast it as a manifestation of a very ancient philosophical issue, first described in classical antiquity by Plato. This treatment of the problem may help medical professionals recognise and address it.

4 Meno's paradox

A helpful formulation of the problem with which the medics described above have been struggling can be found in the *Meno*, which is one of Plato's Socratic dialogues. This formulation of the problem is referred to as Meno's paradox. The Meno is a stylized conversation held largely between two characters—Meno and Socrates. The main subject of this dialogue is the nature of virtue, but at one point during the discussion Meno reaches a point of despair, and argues that it is not possible to investigate the nature of anything at all:

> And how will you inquire into this, Socrates, when you don't know at all what it is? For what sort of thing, from among those you don't know, will you put forward as the thing you are inquiring into? And even if you really encounter it, how will you know if this is the thing that you did not know? (Fine 2014, p. 7).

In response, Socrates immediately reformulates this challenge from Meno into a dilemma:

> I understand the sort of thing you want to say, Meno. Do you see what an eristic argument you're introducing, that it's not possible for someone to inquire either into that which he knows or into that which he doesn't know? For he wouldn't inquire into that which he knows (for he knows it, and there's no need for such

[9] Circular arguments like the ones identified in this paper are quite common in the medical literature. Some other examples can be found in the medical literature on thyroid disease (Gøtzsche 2007, pp. 80–81), giant cell arteritis (Hunder et al. 1990), Takayasu arteritis (Arend et al. 1990), and non-accidental head injury (Högberg et al. 2016). In my view, many such arguments are not the result of researchers accidentally incorporating the index test into the reference standard, which is a simple problem to fix. Rather, these arguments result from trying to directly inspect patients to discover the definitive diagnostic criteria for a disease, and failing to realize that this is not possible.

> a person to inquire); nor into that which he doesn't know (for he doesn't even know what he'll inquire into) (Fine 2014, pp. 7–8).

This challenge by Meno, and the dilemma that Socrates makes out of it, are together referred to as Meno's paradox (Fine 2014). This passage is deemed paradoxical because it appears to show that it is impossible to investigate anything, which is contrary to everyday experience where it seems that this is possible. There is a large philosophical literature discussing this paradox, what Plato meant by it, and possible ways to resolve it (Scott 2006, p. 75). In this paper I read Meno's paradox in a particular way, not because I believe that this is the only correct way to read it, but rather because this reading helps me to highlight a serious epistemic difficulty faced by researchers trying to evaluate diagnostic tests. For the purposes of this paper, I read Plato presenting the problem that if one cannot define something, one cannot recognise the instances of it in the world; and if one cannot do this, one cannot inspect this thing to discover how it should be defined.[10] Applying this passage to a medical context, when researchers do not know how to *identify* the patients that have a disease, it is impossible to discover how to identify these patients *by direct observation of their distinguishing characteristics.*

It is useful to apply Meno's challenge to medicine, especially if the group of patients that have a disease are taken to be the thing that researchers want to investigate. Meno's challenge will raise its head whenever doctors do not know how to identify this group of patients, and seek to discover how to identify them by direct observation. Doctors can find themselves in exactly this situation when they try to measure the accuracy of candidate diagnostic criteria when there is no agreed upon gold or reference standard. Perhaps this is because no-one has suggested a reference standard, or because doctors disagree about which reference standard is best. In any event, the general scheme of research that seeks to measure diagnostic accuracy is always the same, and how Meno's challenge manifests can be understood in these terms.

Researchers seeking to measure diagnostic accuracy always start with a population of patients, some of whom are diseased, and some of whom are not (see Fig. 1). The characteristics that define the diseased group of patients are the definitive diagnostic criteria. As the researchers seek to discover these definitive diagnostic criteria, it follows that they do not know what these are at the start of their study. In order discover these characteristics, researchers may want to directly inspect diseased patients, so

[10] The claim that if one does not know the definition of something then one cannot determine that thing's extension is problematic. Following Locke (1996, p. 185), if something's definition is taken to be its real essence, or what makes the thing the thing that it is, then it may be possible to know the thing's nominal essence, or the set of observable characteristics by which it can be recognised, without knowing its definition. In the Meno, Plato appears to reject this possibility, and scholars have found this problematic. I also agree with Ebrey (2014) that Plato is concerned with discovering "explanatory definitions", or real essences. Here I only take 'definition' to mean the way by which the extension of something can be fixed, or how it can be *recognised*. I follow Fine (2014, p. 73), Irwin (1995, p. 131) and Scott (2006, p. 77) in reading Plato as suggesting that if we cannot recognise something then we cannot fix its extension in order to study it. However, I follow Ebrey (2014) in reading phrase "don't know at all what it is" as not signalling that one needs to be in a "complete mental blank" about something for the paradox to apply, as these scholars suggest (see footnote 17).

that they can be compared to one another and to not-diseased patients.[11] To do this, however, doctors must first be able to identify the diseased and not-diseased patients. So all such research has two stages. In the first stage, patients are sorted into diseased and not-diseased groups. The methods used to do this serve as the gold or reference standard. In the second stage, these patients are inspected to identify distinguishing characteristics, or to evaluate an index test. If researchers do not know how to identify the group of patients with a disease, then they cannot discover distinguishing characteristics by direct observation, because in order to do this they need to identify patients with the disease, and this is precisely what they do not know how to do.

Researchers can, of course, discover further ways of distinguishing diseased and not-diseased patients by directly inspecting them. Given that researchers know how to distinguish these patients, it is possible to directly inspect them to discover other characteristics that can be used to make this distinction. In Fig. 1, researchers use characteristic Y to sort patients into diseased and not-diseased groups. Having done this, researchers can then directly inspect these groups to discover other characteristics that can be used to make the same distinction, such as characteristic Z.

What it is not possible to do is discover that characteristic Y can distinguish between diseased and not-diseased patients by direct inspection. Characteristic Y is deemed to be the most accurate test of all, as it is the test by which the accuracy of other tests are judged. It serves as the gold or reference standard, and might even be considered as the definitive diagnostic criteria for the disease in question. To measure the accuracy of a test is to treat it as an index test—as a test under evaluation—and not as the most accurate test of all. Therefore, the test deemed most accurate of all can never have its accuracy measured. The status of being the most accurate test of all cannot be acquired by measuring diagnostic accuracy through direct observation.

It is impossible to carry out a study that discovers the definitive diagnostic criteria, the gold standard, the reference standard, or the way of diagnosing patients that is most accurate of all, by direct observation. How can researchers study diseased patients when they do not know how to identify them? For which patients, from the population of patients whose disease status they do not know, will they put forward as the group of patients they are inquiring into? Even if they did somehow manage to identify the correct group of patients, how would they know that they had managed to identify this group of patients, seeing as they do not know how to recognise them? Researchers that already know the most accurate way to identify diseased patients have no need of such a study; and researchers who do not know which is the most accurate way to identify patients cannot carry out such a study, for they do not know how to identify the groups of diseased and not-diseased patients in order to study them. Meno's paradox actually manifests in medical practice today.

A problem very like Meno's paradox has been identified in more recent philosophical literature, and it may be valuable to consider the problems faced in medical practice from this perspective as well. Chang (2004, p. 59) has drawn attention to a problem that crops up when trying to measure temperature, which he calls "the problem of

[11] The group of not-diseased patients may be comprised of patients who are entirely free of disease, or to patients free of the particular disease the diseased patients have even though they do have some other disease.

nomic measurement". When trying to construct an accurate thermometer, researchers need to know how a particular thermometric fluid (perhaps mercury or air) expands as its temperature increases. Ideally, a thermometric fluid would expand linearly with increasing temperature, so the same increase in temperature would produce the same amount of expansion at any starting temperature. But in order to measure how a thermometric fluid expands, researchers need to be able to measure temperature of the thermometric fluid, which means they need an accurate thermometer, and they do not have one as this is what they are trying to make. It seems that unless researchers already have an accurate thermometer, they cannot develop an accurate thermometer. This is analogous to the problem in medicine, where researchers cannot develop an accurate reference standard unless they already have an accurate reference standard.

Chang presents this problem formally as a general issue which will crop up anytime researchers try to measure something that cannot be directly observed.[12] He presents the problem of nomic measurement as follows: "We want to measure some quantity X", which in Chang's example is the temperature of some object. "Quantity X is not directly observable, so we infer it from another quantity, Y, which is directly observable". Quantity Y in Chang's example is the volume of some thermometric fluid. Chang notes that if researchers are to make inferences about X using observations of Y, then they need to know how X and Y are related. "For this inference we need a law that expresses X as a function of Y, as follows: X = f(Y)". He argues that this relationship between X and Y cannot be established by direct observation, as X is not directly observable. "The form of this function f cannot be discovered or tested empirically, because that would involve knowing the values of Y and X, and X is the unknown variable we are trying to measure" (Chang 2004, p. 59).

Chang calls this issue the problem of *nomic* measurement because he argues it will apply "Whenever we have a method of measurement that rests on an empirical law" expressing the relationship between two variables (Chang 2004, p. 59). The function notation used by Chang to describe the continuous relationships captured in empirical laws is difficult to apply to the cases I discuss here. The relationship between disease status and diagnostic test results are usually expressed in terms of sensitivity and specificity, which give the probability of patients *with and without* a disease getting a particular test result. This treats disease status as a categorical variable, not as a continuous one.[13] Chang's formulation of the problem of nomic measurement can be modified to reflect this, and written as follows:

1. We want to detect all the patients suffering from a disease, who can collectively be referred to as X (X might be those patients suffering from EMS or fibromyalgia).
2. It is not known how to detect this group of patients, X, so we infer their presence or absence using another quantity, Y, which we know how to detect (Y might be

[12] When Chang (2004) discusses things that cannot be observed directly, he is referring to metaphysical entities that it is not possible to observe directly. Even so, this problem will still manifest in situations where the thing in question is observable, but researchers do not know how to observe it. Consequently, this is what I mean by cannot be directly observed here.

[13] Some diseases, such as hypertension, do lend themselves to Chang's continuous formulation. If researchers choose to treat disease status as a continuous variable, then perhaps Chang's original formulation should be used.

the presence of EMS patterns 1 or 2 in the absence of an alternative explanation for this presentation in the case of EMS, or the presence of widespread pain and 11 out of 18 tender points in the case of fibromyalgia).
3. To make this inference, we need to know the relationship between X and Y.
4. We cannot determine this relationship between X and Y empirically, as to do this we would need to be able to detect both X and Y, and X is the thing we are trying to detect.

When researchers do not know how to identify the group of patients with a disease, it is impossible to show that certain test results correlate with disease status. This is because disease status needs to be known to make this correlation and disease status is unknown. The problem of nomic measurement is a valuable reformulation of Meno's paradox, and may assist researchers in their thinking about the evaluation of diagnostic tests.

Chang also identifies that the problem of nomic measurement can lead to the development of circular arguments in the attempt to justify the accuracy of a method of measurement. As Chang points out, any attempt to demonstrate empirically that a particular relationship between X (the condition we cannot detect directly) and Y (the observations we hope to use to detect X) will result in a circular argument (Chang 2004, p. 89). This is because X cannot be detected without knowing the relationship between X and Y, and the relationship between X and Y cannot be determined without being able to detect X (Chang 2004, p. 89). By assuming that a certain relationship between disease status (X) and diagnostic test result (Y) exist in order to identify diseased patients and carry out their studies, and then claiming to have established this relationship by direct observation of how these things are correlated, researchers often assume what they claim to have shown, and deploy circular arguments.

To see how circular arguments are produced in medical literature that tries to assess the diagnostic accuracy of the gold or reference standard, attend to Figs. 1 and 2 together. In Fig. 1, characteristic Y is used to sort patients into diseased and not-diseased groups. The argument used to do this is presented in Fig. 2 as argument A. Having sorted patients into diseased and not-diseased groups, researchers can inspect these groups to discover distinguishing characteristics in stage 2. As discussed above, researchers are welcome to use their results to support the view that characteristic Z is only present in diseased patients. In Fig. 2, there is nothing wrong with using argument B to follow argument A. However, there is something wrong with using argument C to follow argument A. The conclusion of argument C is that patients with characteristic Y have the disease, and this is taken as a premise in argument A, making the overall argument circular. Any time researchers try to use observations made of diseased and not-diseased patients to support the view that the way they sorted patients into diseased and not-diseased groups is accurate will result in the production of circular arguments. This will happen inevitably when researchers try to directly observe how accurate the gold or reference standard is.

5 Rickets in infants

I will draw attention to one further example of researchers trying to directly observe the characteristics of patients with and without disease. This is important to do, Meno's

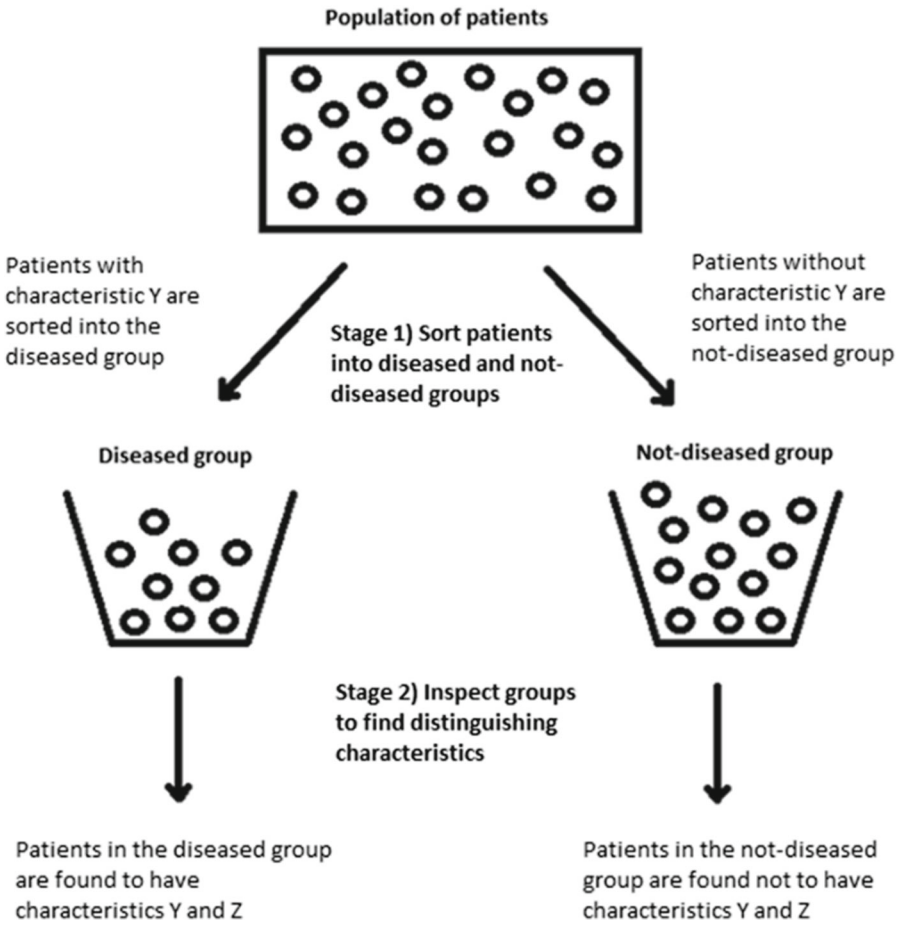

Fig. 1 General scheme for all research measuring diagnostic accuracy

paradox does not only manifest in formal studies of diagnostic accuracy, where the two-stage structure of such research is made explicit. It also manifests in less formal arguments presented in discussions of how to diagnose disease, and this should be highlighted.

My last case study is an article by Drs. Slovis and Chapman (2008b), which is concerned with the diagnosis of rickets in infant children. The diagnosis of rickets in infants is important in both legal and medical contexts. Parents or carers of infants are sometimes accused of assault if the infants under their care are found to have unexplained fractures. Rickets has been suggested as a possible explanation for these fractures, and therefore as an alternative explanation to abuse.

Slovis and Chapman are the editors of a well-respected medical journal, *Pediatric Radiology*, and were responding in their article to another article in the same edition of this journal that argued that rickets should be considered as an alternative diagnosis for patients with unexplained fractures (Keller and Barnes 2008). In this other article,

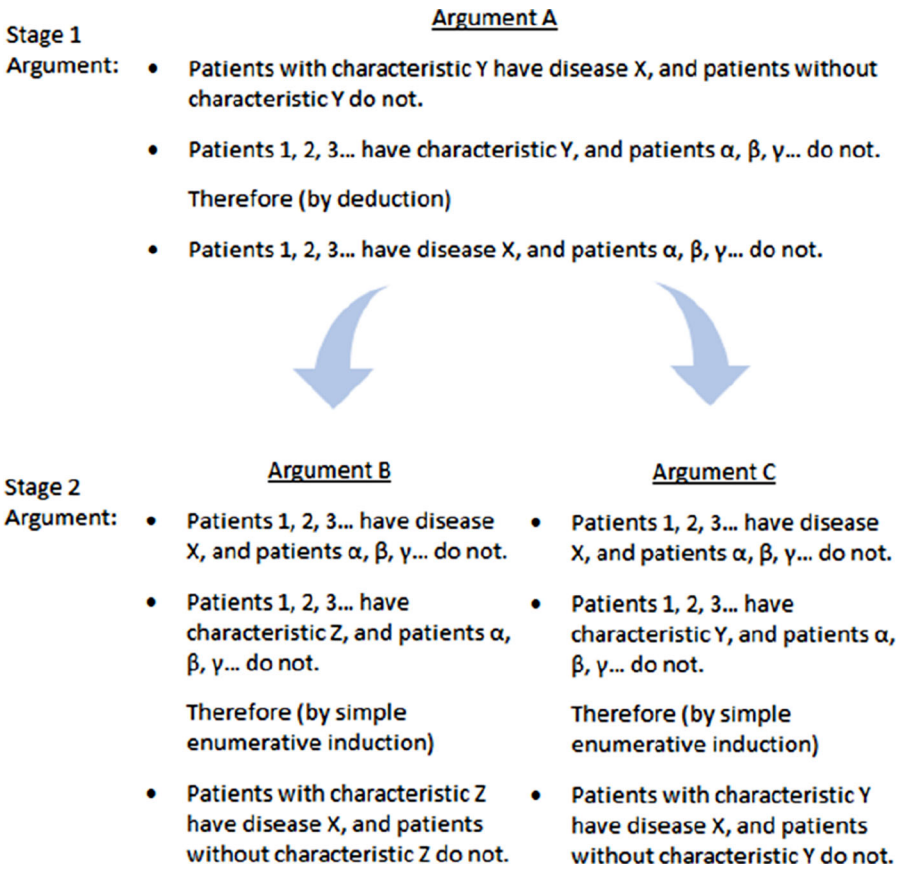

Fig. 2 Arguments that might be deployed in studies of diagnostic accuracy. Argument A is used to identify patients with disease X in stage 1. Argument A might be followed by either argument B or C in stage 2. In argument B, an index test (Z) is evaluated, which is fine. In argument C, the reference standard (Y) is evaluated, producing a circular argument where the conclusion of argument C is the same as the first premise of argument A

Drs. Keller and Barnes (2008) presented four cases with unexplained fractures, which they argued were cases of rickets. They did this even though these cases did not have what are considered to be the classic signs on X-ray examination that are normally used to diagnose rickets. The classic signs of rickets are taken to be changes to a region at the ends of long bones in growing children called the "growth plate" or "physis" (Slovis and Chapman 2008b). Keller and Barnes (2008) argued that these four patients had rickets because of evidence that either the infants or their mothers were vitamin D deficient, and because of other radiographic abnormalities that these infants displayed. Overall, they argued that vitamin D deficiency rickets can produce imaging abnormalities that resemble traumatic fractures without the classic radiographic signs of rickets, and therefore that rickets in the young infant can mimic abuse.

Slovis and Chapman (2008a), as editors of the journal, commented that this is a controversial position. Nevertheless, they encouraged readers of the journal to consider

the evidence that informs this issue carefully, and to draw conclusions based on this data for themselves:

> We believe that it is one of the responsibilities of a medical journal to publish articles that present data that force us to rethink our preconceived notions. We believe it is important that all pediatric radiologists understand this issue, as we play a focal role in the diagnosis of child abuse (Slovis and Chapman 2008a).

Slovis and Chapman (2008b) began their own article by reaffirming that their goal was to entertain the question of whether or not there are patients with rickets that do not have the classic radiographic signs of rickets:

> What is the evidence for fragility of bones in children with insufficient levels of vitamin D and even in those with deficiency levels if the radiographs are normal, that is, when there is no radiographic evidence of rickets? (Slovis and Chapman 2008b, p. 1221).

To address this question, Slovis and Chapman (2008b) chose what may be considered a straightforward, simple and pragmatic methodology. They chose to directly inspect the cases presented by Keller and Barnes (2008) to see if there were any cases of rickets without the classic radiographic signs amongst them. The particular patients in question here were the four cases discussed by Keller and Barnes (2008)—cases 1, 2, 3, and 4. But before Slovis and Chapman (2008b) could look and see whether any of these patients had rickets without the classic radiographic signs of rickets, they needed to determine which of these patients actually had rickets. To do this they used a commonplace definition of rickets to decide if a patient had rickets or not:

> The definition of rickets is "an interruption in the development and mineralization of the growth plate of bone, with radiographic abnormalities". Merely having insufficiency/deficiency of vitamin D levels in the blood does not constitute rickets. It is, therefore, incumbent to show radiographic changes in the 30–50% of infants and children with low vitamin D to claim that they have rickets (Slovis and Chapman 2008b, p. 1221).

For a patient to have rickets, Slovis and Chapman (2008b) said that the patient must have certain changes to the growth plate of their bones. In other words, they must have the classic radiographic changes associated with rickets. Slovis and Chapman (2008b) argued that none of these cases had rickets because none of them had the classic radiographic signs of rickets:

> It is apparent in all the images of Keller and Barnes that the epiphysis and metaphysis are not separated and the physis is normal. There is no cupping and fraying. By definition, radiological rickets is not present in these images (Slovis and Chapman 2008b, p. 1222).[14]

[14] Slovis and Chapman (2008b) use a variety of different terms to describe rickets and the related diagnostic observations. For instance, they sometimes refer to "radiographic abnormalities", to "radiographic changes", and to "radiographic findings" (2008b, p. 1221). They also refer to "radiographic rickets" and "radiological rickets", as well as simply to "rickets" (2008b, p. 1222). This opens the possibility that they are referring to different types of rickets with different sets of diagnostic criteria. In my view, abnormalities/changes/findings

As none of the cases presented by Keller and Barnes (2008) were deemed to have rickets, Slovis and Chapman (2008b) argued that none of these cases provided evidence that there are patients with unexplained fractures who are cases of rickets who do not have the classic radiographic signs of rickets. "For these reasons and because of the other data described, we find that the connection made by Keller and Barnes between "rickets" and fractures they consider to be similar in appearance to those seen in child abuse is not based on any scientific data" (Slovis and Chapman 2008b). In the absence of evidence to the contrary, they concluded that a diagnosis of rickets should not be made without radiographic evidence of changes to the growth plate:

> The work-up of child abuse considers a differential diagnosis including rickets but, unless there is reasonable evidence of rachitic bone disease, there is no scientific basis for confusing vitamin D insufficiency/deficiency alone with child abuse (Slovis and Chapman 2008b, p. 1224).

According to this definition, it is *necessary* that a patient have certain changes to the growth plate to be considered as a case of rickets. The reason Slovis and Chapman (2008b) found that none of the cases presented by Keller and Barnes (2008) had rickets was *because* these cases showed none of these radiographic changes in the region of the growth plate. So the conclusion that patients do not have rickets unless they have these radiographic changes to their growth plate is supported by the assumption that patients do not have rickets unless they have these radiographic changes to the growth plate. Slovis and Chapman (2008b) assumed what they claimed to have shown, and deployed a circular argument.[15]

Again, I argue that this circular argument is connected to the problem of nomic measurement.[16] Slovis and Chapman (2008b) sought to provide evidence that rickets (X) does not occur without radiographic changes at the growth plate (Y). But the

Footnote 14 continued
are used interchangeably. I read "radiographic" and "radiological" rickets refer to rickets that has been properly diagnosed, or at least to a necessary component of the proper diagnosis of rickets.

[15] Slovis and Chapman's (2008b) argument can be written out formally. As in Figs. 1 and 2, this argument is made in two stages. The first stage determines whether or not the four cases cited by Keller and Barnes (2008) are cases of rickets:

- All cases of rickets show the classic radiographic signs associated with rickets.
- Cases 1, 2, 3, and 4 do not have the classic radiographic signs of rickets.
  Therefore (by deduction)
- Cases 1, 2, 3, and 4 do not have rickets.

The second stage determines whether there are any cases of rickets without the classic radiographic signs of rickets amongst these cases:

- Cases 1, 2, 3, and 4 do not have rickets.
- Cases 1, 2, 3, and 4 do not have the classic radiographic signs of rickets
  Therefore (by simple enumerative induction)
- There are no cases of rickets that do not have the classic radiographic signs associated with rickets.

The conclusion of the second stage argument is logically equivalent to the first premise of the first stage argument. The overall argument is therefore circular.

[16] Slovis and Chapman (2008b) stipulate that, by definition, all patients with rickets have the classic radiographic signs of rickets. They may not view this association of rickets and radiographic signs as law of nature which they trust, but rather as part of the meaning of rickets. Chang's problem of nomic measurement,

relationship between rickets and radiographic changes at the growth plate cannot be determined by observation unless the presence of rickets can be determined, and Slovis and Chapman (2008b) determine the presence of rickets using knowledge of the relationship between rickets and radiographic changes at the growth plate. By seeking to evaluate what they take to be the definitive diagnostic criteria for rickets, Slovis and Chapman (2008b) accept that they do not know how to identify patients with rickets (at least for the purposes of their paper). By trying to make this evaluation by directly inspecting patients to see whether there are any patients with rickets who do not satisfy these criteria, Slovis and Chapman (2008b) take on an impossible task. If researchers do not know how to identify the group of patients with rickets, then they cannot directly observe what the distinguishing characteristics of these patients are, because in order to do this they need to identify patients with rickets, and this is precisely what they do not know how to do.

In 2013, 5 years after Keller and Barnes (2008) and Slovis and Chapman (2008b) had presented their arguments, Strouse (2013, p. 1423) revisited this matter in the same journal. Strouse (2013) argued that Keller and Barnes' (2008) arguments were refuted by Slovis and Chapman's (2008b) arguments, as well as by subsequent work:

> Perhaps better stated, the hypotheses of "Keller & Barnes" have been disproved. Whether in legal proceedings or in medical literature, it is inexcusable and inappropriate to cite "Keller & Barnes," particularly without simultaneously citing the accompanying commentary by Slovis and Chapman and the subsequent commentary by Feldman (Strouse 2013, p. 1424).

This shows that, even in 2013, many researchers still had not recognised that Slovis and Chapman's (2008b) argument is circular. That Slovis and Chapman (2008b) have set themselves an impossible task has not been realised either.

Slovis and Chapman (2008b) and Keller and Barnes (2008) have a fundamental disagreement about how to identify patients with rickets. Even though they may agree about which patients have rickets in many cases, they still disagree in these few cases.[17] Such a disagreement cannot be resolved by directly inspecting patients with rickets, as this requires that these parties agree about which patients have rickets, and this is what they are arguing about. Meno's paradox manifests even when researchers are evaluating diagnostic criteria informally, without explicitly measuring sensitivities and specificities of diagnostic tests.

6 There are other ways to evaluate reference standards

I have argued that it is impossible to discover how to distinguish diseased and not-diseased patients by directly inspecting these patients, in an effort to correlate characteristics with disease status. It would be remiss to leave the reader with the

Footnote 16 continued
which refers to laws of nature that are trusted enough to use to make measurements, may not strictly apply here. Nevertheless, Chang's insight still provides a valuable way to think about the rickets case.

[17] This shows that researchers do not have to be in a "complete mental blank" about something for the problems associated with Meno's paradox to manifest (see footnote 10).

impression that I believe it to be impossible to investigate how to distinguish diseased from not-diseased patients. It is useful to close with a few comments on alternative ways of investigating how to identify diseased patients.

In the introduction, I drew attention to Reitsma et al. (2009), who review the different methods used by medical researchers to evaluate the accuracy of diagnostic tests in the absence of a gold or reference standard. All of these methods deserve greater philosophical attention. Reitsma et al. (2009) argue that even though these methods might be useful in different circumstances, none of them provide a generally reliable method of measuring diagnostic accuracy in the absence of a reference standard.

In addition to these methods of measuring diagnostic accuracy, Reitsma et al. (2009) suggest that the effort to identify accurate diagnostic tests might be abandoned altogether. "If none of these methods for repairing standard imperfections seem appropriate, the diagnostic accuracy paradigm may be abandoned in favor of alternative methods for evaluating tests" (Reitsma et al. 2009, p. 903). This is quite a dramatic shift in thinking about what diagnostic tests are supposed to do. Medical researchers, reflecting on this different philosophical attitude, describe it as a shift from being essentialist,[18] and worrying about finding tests that capture the correct classification of patients in a deep ontological sense, to being consequentialist, and worrying about how adopting one test rather than another affects patient outcomes (Patrick Bossuyt, co-author of STARD and QUADAS, personal communication). Instead of worrying about whether a test is accurate, Reitsma et al. (2009) join others in pointing out that that researchers could instead worry about whether a test is useful (see also Haynes and You 2009; Newman and Kohn 2009, p. 7).[19] In this context utility refers to the ability of the test to select patients who have some property that is of interest. Perhaps they generally respond better to a particular therapy, or perhaps it is possible to offer an accurate prognosis for these patients. Utility is complex to assess, as the usefulness of a test will depend on the clinical circumstances in which it is applied. This can lead to pluralistic classificatory practices, where doctors argue that "There is no single best way to classify people into those with different diagnoses; the optimal classification scheme depends on the purpose for making the diagnosis" (Newman and Kohn 2009, p. 8).[20]

However, as Reitsma et al. (2009) correctly point out, that even this approach is not unproblematic. They argue (in a manner commensurate with the Duhem–Quine thesis

[18] As noted in the introduction, Meno's paradox can still cause problems when trying to measure diagnostic accuracy even when diseases are understood as subjective things, which do not have essences. Some researchers have nevertheless used the language of essentialism to describe their concerns about the evaluation of diagnostic tests.

[19] Both Haynes and You (2009) and Newman and Kohn (2009) recommend that the accuracy of a test be measured before its usefulness is considered, so that resources are not wasted on inaccurate tests "Again if a test is not accurate, you can stop; it cannot be useful" (Newman and Kohn 2009, p. 7). This makes the measurement of accuracy an important part of the demonstration of usefulness. Reitsma et al. (2009) are more radical than these authors in suggesting that researchers focus on usefulness rather than accuracy.

[20] Drawing on the work of Heinrik Wulff (1976) and Alvan Feinstein (1967), H. Tristram Engelhardt has long been arguing along these lines. "What I have been suggesting here is a proposal for constructing various alternate typologies of disease, aimed at facilitating different sorts of clinical decision-making…It is in this sense that typologies of disease are not true or false in any straightforward fashion but rather are more or less useful in the conduct of clinical medicine" (Engelhardt 1985, p. 67).

in the philosophy of science) that if the patients selected by a diagnostic practice do not have the expected clinical outcomes, then there will always be a number of ways in which this conflict with experience can be explained: "Whenever the index test results fail to show the hypothesized network of associations, more than one conclusion is possible—the index test has low validity, the theory about the target condition is not correct, or both" (Reitsma et al. 2009, p. 804). Given this, even focusing on the clinical outcomes of adopting different diagnostic practices does not fully determine which diagnostic practices should be adopted. It seems that there are no uncontroversial solutions to the problem of how to evaluate diagnostic practices in the absence of an agreed upon reference standard in the wider literature.[21] The utility of all these different potential ways of evaluating diagnostic practices is an important area of ongoing research in medical practice, and one to which philosophers of science and medicine should be able to contribute.

Another alternative approach is not necessarily to abandon the paradigm of diagnostic accuracy, but rather to change the method of inference used to assess diagnostic accuracy. As discussed above, diagnostic accuracy is usually assessed by correlating diagnostic test results with disease status. Finding that certain test results correlate well with disease status in a study, and perhaps finding that this correlation is reproduced in other studies, is taken to indicate that this correlation will hold generally in the future. However, researchers are not forced to investigate disease using such inductive inferences alone—*abduction* may also be useful.[22] Abduction is a method of inference that is used to explain observations that at first do not make sense to researchers (Lipton 2004; Hanson 1958, p. 86). "Its governing idea is that explanatory considerations are a guide to inference, that scientists infer from the available evidence to the hypothesis which would, if correct, best explain that evidence" (Lipton 2008, p. 193).

Abduction is used by medical researchers, even if they do not recognise this explicitly. In addition to the argument discussed above, Slovis and Chapman (2008b, p. 1223) also argue using abduction. They argue that it is surprising that there are so few patients with unexplained fractures, given that there is an epidemic of infants with low vitamin D levels. This would be explained if low vitamin D does not cause increased skeletal fragility and unexplained fractures. So, there is reason to believe that the epi-

[21] Reitsma et al. (2009) do not consider the possibility of using pathophysiological knowledge to evaluate reference standards. Other researchers, however, do. For example, Knottnerus and Muris (2009, p. 55) argue that it is important to use pathophysiological knowledge to produce new and better reference standards, because evidence of correlation alone will never be able to demonstrate that a new reference standard is more accurate than the old one. "Therefore, pathophysiological expertise should be involved in the evaluation of diagnostic accuracy" (Reitsma et al. 2009, p. 55). This attitude results in something of a paradox, as these proponents of evidence-based medicine found their knowledge of diagnosis on pathophysiological expertise. As knowledge of whether a treatment works to cure a particular disease is itself founded on knowledge of diagnosis, this means that evidence of the efficacy of treatment is also founded on pathophysiological expertise. This runs counter to the tenets of evidence-based medicine, which downplays the importance of pathophysiological expertise (Clarke et al. 2014; Solomon 2015).

[22] Induction is sometimes taken to refer to any ampliative form of inference, but here I use induction to refer to induction by enumeration (Hawthorne xxxx). The term 'abduction' is used in a variety of different ways in philosophical literature. In particular, some commentators want to distinguish between 'abduction' as an inference that suggest new hypotheses, and 'inference to the best explanation' an inference that gives reason to accept an existing hypothesis in preference to others (Douven xxxx; Mcauliffe 2015; Campos 2011). I don't make use of this distinction here.

demic of low vitamin D levels does not cause increased levels of skeletal fragility and unexplained fractures. The diagnosis of EMS also appears to have been originally suggested as the result of an abduction, as researchers tried to make sense of a number of patients who presented with strange and unexplained symptoms (Bulpitt et al. 1990; Hertzman et al. 1990). Research into the historical processes by which these patterns were identified may be useful in contemporary conversations about how to diagnose disease. Abduction has been identified as an important method of inference in medical discovery and diagnostics (Johansson and Lynø 2008, pp. 121–125; Walton 2005, p. 175). As Sami Paavola and Hakkarainen (2005) have argued, abduction may be a way to resolve the problems presented by Meno's paradox.[23]

7 Conclusion

The measurement of diagnostic accuracy is an important aspect of the evaluation of diagnostic tests. Diagnostic accuracy is measured by comparing the performance of the observation or test under evaluation against a gold or reference standard, which is the test deemed to be the most accurate of all. This amounts to correlating the presence and absence of an observation or test result with the disease status of patients. Research measuring diagnostic accuracy always has two stages. In the first, the population of patients being studied is divided into diseased and not-diseased groups. In the second, these groups are inspected to find characteristics that can be used to distinguish diseased and not-diseased patients.

Sometimes, however, doctors are unsure about how to identify the patients who actually have the disease in question, and there is debate about what set of diagnostic criteria should be used as the gold standard. In these circumstances, researchers are presented with an ancient philosophical problem. This was first articulated in Plato's *Meno*, and has been discussed again more recently as the problem of nomic measurement (Chang 2004). If researchers do not know how to identify the group of patients with a disease, then they cannot directly observe what the distinguishing characteristics of these patients are, because in order to do this they need to identify patients with the disease, and this is precisely what they do not know how to do. If disease status is unknown, then it is impossible to correlate test results with disease status. Even so, I have provided three examples of medical researchers that try to do just this, and produce circular arguments as a result; drawn from the medial literature on EMS, fibromyalgia, and rickets. That it is impossible to establish what should serve as definitive diagnostic criteria by assessing diagnostic accuracy needs to be more widely recognised.

Acknowledgements I would like to thank University of Exeter Medical Students who have taken the *Thinking critically about diagnosis* Medical Humanities Student Selected Unit, where the issues explored in this paper have been extensively discussed. In particular, I want to thank Thomas Bennett and Jess Chan

[23] I will say no more here about whether the many other ways scholars have sought to resolve Meno's paradox (see references in footnote 6) are successful in this medical context; or indeed about whether Chang's (2004) account of how researchers developing thermometers resolved the problem of nomic measurement would be applicable to medicine. I only wish show that these problems are not insurmountable.

for drawing my attention to the fibromyalgia case. I am grateful to Lara Keuck and Chris Hyde for their encouragement, and to Jonathan Davies for introducing me to Meno's paradox. Research and open access publication funded by the University of Exeter.

References

Arend, W. P., Michel, B. A., Bloch, D. A., Hunder, G. G., Calabrese, L. H., Edworthy, S. M., et al. (1990). The American College of Rheumatology 1990 criteria for the classification of Takayasu arteritis. *Arthritis & Rheumatology*, *33*(8), 1129–1134.

Barker, K. (2005). *The fibromyalgia story: Medical authority and women's worlds of pain*. Philadelphia: Temple University Press.

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., et al. (2003). The STARD statement for reporting studies of diagnostic accuracy: Explanation and elaboration. *Clinical Chemistry*, *49*(1), 7–18.

Bossuyt, P. M., Reitsma, B. J., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L., et al. (2015). STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. *Radiology*, *277*(3), 826–832.

Bulpitt, K. J., Verity, M. A., Clements, P. J., & Paulus, H. E. (1990). Association of l-tryptophan and an illness resembling eosinophilic fasciitis: Clinical and histopathologic findings in four patients with eosinophilia-myalgia syndrome. *Arthritis & Rheumatology*, *33*(7), 918–929.

Campos, D. (2011). On the distinction between Peirce's abduction and Lipton's Inference to the best explanation. *Synthese*, *180*(3), 419–442.

Cardoso, J. R., Pereira, L. M., Iversen, M. D., & Ramos, A. L. (2014). What is gold standard and what is ground truth? *Dental Press Journal of Orthodontics*, *19*(5), 27–30.

Chang, H. (2004). *Inventing temperature: Measurement and scientific progress*. Oxford: Oxford University Press.

Clarke, B., Gillies, D., Illari, P., Russo, F., & Williamson, J. (2014). Mechanisms and the evidence hierarchy. *Topoi*, *33*(2), 339–360.

Cohen, M. L., & Quintner, J. L. (1993). Fibromyalgia syndrome, a problem of tautology. *The Lancet*, *342*(8876), 906–909.

Douven, I. (2017). Abduction. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Summer 2017 Edition). https://plato.stanford.edu/archives/sum2017/entries/abduction/.

Duggan, P. F. (1992). Time to abolish "gold standard". *British Medical Journal*, *304*, 1568–9.

Ebrey, D. (2014). Meno's paradox in context. *British Journal for the History of Philosophy*, *22*(1), 4–24.

Engelhardt, H. T. (1985). Typologies of disease: Nosologies revisited. In K. F. Schaffner (Ed.), *Logic of Discovery and diagnosis in medicine* (pp. 56–71). Berkeley: University of California Press.

Ereshefsky, M. (2009). Defining 'health' and 'disease'. *Studies in the History of the Biological and Biomedical Sciences*, *40*, 221–227.

Feinstein, A. R. (1967). *Clinical judgment*. Baltimore: Lippincott Williams & Wilkins.

Fine, G. (2014). *The possibility of inquiry: Meno's paradox from socrates to sextus*. Oxford: Oxford University Press.

Fontela, P. S., Pant Pai, N., Schiller, I., Dendukuri, N., Ramsay, A., & Pai, M. (2009). Quality and reporting of diagnostic accuracy studies in TB, HIV and malaria: Evaluation using QUADAS and STARD standards. *PLoS ONE*, *4*(11), e7753. https://doi.org/10.1371/journal.pone.0007753.

Gøtzsche, P. C. (2007). *Rational diagnosis and treatment: Evidence-based clinical decision-making*. Chichester: Wiley.

Hanson, N. R. (1958). *Patterns of discovery*. Cambridge: Cambridge University Press.

Hawthorne, J. (2012). Inductive logic. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Spring 2017 Edition). https://plato.stanford.edu/archives/spr2017/entries/logic-inductive/.

Haynes, B. R., & You, J. J. (2009). The architecture of diagnostic research. In J. A. Knottnerus & F. Buntinx (Eds.), *The evidence base of clinical diagnosis: Theory and methods of diagnostic research* (pp. 20–41). Chichester: Wiley-Blackwall.

Hertzman, P. A., Blevins, W. L., Mayer, J., Greenfield, B., Ting, M., & Gleich, G. J. (1990). Association of the eosinophilia–myalgia syndrome with the ingestion of tryptophan. *New England Journal of Medicine*, *322*, 869–873.

Hertzman, P. A., Clauw, D. J., Duffy, J., Medsger, T. A, Jr., & Feinstein, A. R. (2001). Rigorous new approach to constructing a gold standard for validating new diagnostic criteria, as exemplified by the eosinophilia–myalgia syndrome. *Archives of Internal Medicine*, *161*(19), 2301–2306.

Högberg, G., Colville-Ebeling, B., Högberg, U., & Aspelin, P. (2016). Circularity bias in abusive head trauma studies could be diminished with a new ranking scale. *Egyptian Journal of Forensic Sciences*, *6*(1), 6–10.

Hunder, G. G., Bloch, D. A., Michel, B. A., Stevens, M. B., Arend, W. P., Calabrese, L. H., et al. (1990). The American College of Rheumatology 1990 criteria for the classification of giant cell arteritis. *Arthritis & Rheumatology*, *33*(8), 1122–1128.

Irwin, T. (1995). *Plato's ethics*. Oxford: Oxford University Press.

Johansson, I., & Lynø, N. (2008). *Medicine & philosophy: A twenty-first century introduction*. Lancaster: Walter de Gruyter.

Kahn, R. L. (1942). *Serology in syphilis control, principles of sensitivity and specificity*. Baltimore: The Williams & Wilkins Company.

Keller, K. A., & Barnes, P. D. (2008). Rickets vs. abuse: A national and international epidemic. *Pediatric Radiology*, *38*(11), 1210–1216.

Kennedy, A. G. (2016). Evaluating diagnostic tests. *Journal of Evaluation in Clinical Practice*, *22*(4), 575–579.

Knottnerus, J. A., & Buntinx, F. (2009). *The evidence base of clinical diagnosis: Theory and methods of diagnostic research*. Chichester: Wiley-Blackwall.

Knottnerus, J. A., Buntinx, F., & van Weel, C. (2009). General introduction: Evaluation of diagnostic procedures. In J. A. Knottnerus & F. Buntinx (Eds.), *The evidence base of clinical diagnosis: Theory and methods of diagnostic research* (pp. 1–19). Chichester: Wiley-Blackwall.

Knottnerus, J. A., & Muris, J. W. (2009). Assessment of the accuracy of diagnostic tests: The cross-sectional study. In J. A. Knottnerus & F. Buntinx (Eds.), *The evidence base of clinical diagnosis: Theory and methods of diagnostic research* (pp. 42–62). Chichester: Wiley-Blackwall.

Lipton, P. (2004). *Inference to the best explanation* (2nd ed.). London: Routledge.

Lipton, P. (2008). Inference to the best explanation. In S. Psillos & M. Curd (Eds.), *The routledge companion to the philosophy of science* (pp. 193–202). London: Routledge.

Locke, J. (1996) [1689]. *An essay concerning human understanding* (abridged and edited by K.P. Winkler). Cambridge: Hackett Publishing Company.

Mackenzie, J. (1916). *Principles of diagnosis and treatment in heart affectations*. London: Frowde, Hodder & Stoughton.

Mcauliffe, W. H. (2015). How did abduction get confused with inference to the best explanation? Transactions of the Charles S. *Peirce Society: A Quarterly Journal in American Philosophy*, *51*(3), 300–319.

Mebius, A., Kennedy, A. G., & Howick, J. (2016). Research gaps in the philosophy of evidence-based medicine. *Philosophy Compass*, *11*(11), 757–771.

Newman, T. B., & Kohn, M. A. (2009). *Evidence-based diagnosis*. Cambridge: Cambridge University Press.

Okeh, U. M., & Okoro, C. N. (2012). Evaluating measures of indicators of diagnostic test performance: Fundamental meanings and formulars. *Journal of Biometrics and Biostatistics*,. https://doi.org/10.4172/2155-6180.1000132.

Paavola, S., & Hakkarainen, K. (2005). Three abductive solutions to the Meno paradox-with instinct, inference, and distributed cognition. *Studies in Philosophy and Education*, *24*(3–4), 235–253.

Reitsma, J. B., Rutjes, A. W., Khan, K. S., Coomarasamy, A., & Bossuyt, P. M. (2009). A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. *Journal of Clinical Epidemiology*, *62*(8), 797–806.

Scott, D. (2006). *Plato's Meno*. Cambridge: Cambridge University Press.

Simon, J. (2017). Realism and constructivism in medicine. In M. Solomon, J. R. Simon, & H. Kincaid (Eds.), *The Routledge companion to philosophy of medicine* (pp. 90–100). New York: Taylor & Francis.

Slovis, T. L., & Chapman, S. (2008a). Vitamin D insufficiency/deficiency—A conundrum. *Pediatric Radiology*, *38*(11), 1153.

Slovis, T. L., & Chapman, S. (2008b). Evaluating the data concerning vitamin D insufficiency/deficiency and child abuse. *Pediatric Radiology*, *38*(11), 1221–1224.
Solomon, M. (2015). *Making medical knowledge*. Oxford: Oxford University Press.
Strouse, P. J. (2013). 'Keller & Barnes' after 5 years—Still inadmissible as evidence. *Pediatric Radiology*, *43*(11), 1423–1424.
Versi, E. (1992). "Gold standard" is an appropriate term. *British Medical Journal*, *305*, 187.
Walton, D. (2005). *Abductive reasoning*. Tuscaloosa: University of Alabama Press.
Weinstein, S., Obuchowski, N. A., & Lieber, M. L. (2005). Clinical evaluation of diagnostic tests. *American Journal of Roentgenology*, *184*(1), 14–19.
Wessely, S., & Hotopf, M. (1999). Is fibromyalgia a distinct clinical entity? Historical and epidemiological evidence. *Best Practice & Research Clinical Rheumatology*, *13*(3), 427–436.
Whiting, P. F., Rutjes, A. W., Reitsma, J. B., Bossuyt, P. M., & Kleijnen, J. (2003). The development of QUADAS: A tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology*, *3*(1), 25.
Whiting, P. F., Rutjes, A. W. S., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., et al. (2011). QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine*, *155*(8), 529–536.
Wolfe, F., Smythe, H. A., Yunus, M. B., Bennett, R. M., Bombardier, C., Goldenberg, D. L., et al. (1990). The American College of Rheumatology 1990 criteria for the classification of fibromyalgia. *Arthritis & Rheumatology*, *33*(2), 160–172.
Worster, A., & Carpenter, C. (2008). Incorporation bias in studies of diagnostic tests: How to avoid being biased about bias. *CJEM*, *10*(02), 174–175.
Wulff, H. R. (1976). *Rational diagnosis and treatment*. Oxford: Blackwell.