

## Preface

This book was started in a pub. After meeting to discuss the intricacies of health technology assessment and medical test evaluation at the University of Exeter Medical School, members of the Exeter Test Club would retire to the Mount Radford Inn for a drink. Discussions of test evaluation would continue in a less formal but more playful mode. It is through many hours of pub discussion with Chris Hyde and Patrick Bossuyt that I was made aware of the considerable and long-standing problems that researchers working on test evaluation face. Advice from Chris and Patrick was invaluable when writing this book, and I am forever in their debt. Wide-ranging discussions were slowly distilled into two questions: Why is the ability of a test to detect specific disease entities widely seen as essential if the test is to be useful in medical practice? Why are the two main indices to evaluate diagnostic accuracy, sensitivity and specificity, commonly believed to be constants, or only to vary in a few circumstances such as with the severity of disease? This book is an effort to articulate and address those problems.

Intriguingly, both Chris and Patrick framed their concerns historically. As the above questions indicate, they held that sensitivity and specificity were highly variable, and did not think that a test needed to be able to accurately detect a specific disease entity for it to be useful. They saw widespread beliefs to the contrary as problems for present-day medical practice. However, they were not content to confine their investigation of these problems to the present. They wanted to know how these widespread beliefs about test evaluation had come to be accepted *historically*. They had an intuition that understanding the historical development of these views would help test evaluators in the present day to do their work. I shared the intuition that the history of medicine could be important for medical practice in the present and was happy to investigate the historical development of these widespread beliefs with this aim in mind.

This investigation led to a publication with Chris and Patrick ‘On the origin of sensitivity and specificity.’<sup>1</sup> This work linked the origins of sensitivity and specificity as

---

<sup>1</sup> Binney, Hyde, and Bossuyt 2021

indices of test performance to the development of serology in the early twentieth century, especially with the diagnosis of syphilis and the Wassermann reaction. This came as something of a surprise to me. In my own historical and philosophical work, I draw heavily from the early twentieth-century Polish philosopher and physician Ludwik Fleck. The Wassermann reaction and its relationship to syphilis is the case study of Fleck's main work. Fleck highlighted that the Wassermann reaction was not immunologically specific. However, it was understood as specific in another sense, as tending to return positive tests only in cases of syphilis. Although the Wassermann reaction was not immunologically specific, it was understood as diagnostically specific. This is the birth of specificity in its diagnostic sense. This work on the history of sensitivity and specificity saw the happy confluence of two topics dear to me, the evaluation of medical tests and Fleck's philosophical work.

Whilst working on the origin of sensitivity and specificity, I was fortunate enough to be employed on the project 'Health and Disease as Practical Concepts' at the Erasmus Medical Center, Rotterdam in the Netherlands, headed by Maartje Schermer. This project sought to develop a pragmatic approach to concepts of health and disease, and aimed to address problems that emerge in medical practice related to conceptions of health and disease. This work prompted me to look at problems of test evaluation in a new way. Instead of thinking about test evaluation as an *epistemic* problem, concerned with producing knowledge of how good tests are, I learned to think about test evaluation as an *ontological* problem, concerned with intuitions about the nature of disease. Ontology is a slightly obscure philosophical term for the study of being, addressing questions of what it is to be and what things are. What are diseases? Are there specific disease entities? What is the character of disease? These are ontological questions. Ontological assumptions about whether patients with the same disease are highly homogeneous or heterogeneous are central to epistemic intuitions about how best to evaluate medical tests. Such questions may seem too remote from medical practice for their answers to be of use to practical medical people, but medical practice involves such assumptions whether medical practitioners consider them closely or not. The health and disease project involved close collaboration between medical practice, historians and philosophers of medicine. Together we developed a working method in which we identify and articulate problems in medical practice associated with concepts of health and disease, carry out historical research to understand how these problems emerged, reflect on how the historical work changed our understanding of the problem in the present day, and finally consider how to respond to the problem. This book broadly follows this method, which we call 'history and philosophy for medicine'. I am grateful to Maartje Schermer, Timo Bolt, Rik van der Linden and Floor Haalboom for their help and support.

During my time at the Erasmus MC, I was also privileged to collaborate with the project 'Practices of Validation in the Biomedical Sciences' at the Max Planck Institute for the History of Science in Berlin, headed by Lara Keuck. I am grateful to Lara and

Alfie Freeborn for looking over my scribbles on test evaluation, and for encouraging me to develop them further. Indeed, it was Lara who sent me the advertisement for the job at the Erasmus MC, so I am doubly indebted to her.

Finally, I will always be grateful to Heiner Fangerau for agreeing to collaborate on the Marie Skłodowska-Curie Postdoctoral Fellowship 'Is This a Good Medical Test?', which produced this book. Working at the Institut für Geschichte, Theorie und Ethik der Medizin, Heinrich Heine University Düsseldorf, has forged links with a network of collaborators from across Europe working on how to improve test evaluation, putting history and philosophy for medicine into practice.

The purpose of this book is not to propose or defend any particular way of evaluating tests. Rather, it is to highlight the basic philosophical assumptions incorporated into ways of evaluating medical tests. Whatever indices and procedures are chosen to evaluate a medical test, they will incorporate basic philosophical assumptions about the character of disease and the purpose of testing. I want to encourage physicians and researchers to consider these assumptions more carefully when evaluating medical tests. I also want to promote the practice of using medical history to reflect on medical practice in the present day. The historical material discussed in this book provides several examples of times when deeper historical knowledge would have benefitted medical practice, and I hope encourages readers to incorporate the history of medicine into their medical practice.

## Introduction

Testing is central to medical practice. When thinking of medical tests, blood tests or other laboratory tests may come most readily to mind. However, medical tests can be understood to consist of any and all possible observations, either on their own or in conjunction with others, that are made of patients. Such tests might include symptoms, signs, observations made with clinical equipment (such as the stethoscope or thermometer), blood tests, urine analysis, microscopy, radiographs and other imaging procedures, genetic tests, electrocardiographs, histological tests, and post-mortem examinations. Such tests can be done for many purposes. When thinking about the purpose of medical tests, *diagnostic* tests may come most readily to mind. However, diagnosis is not the only purpose of testing. Medical tests are carried out for many reasons, including to direct therapy, to monitor response to therapy, to screen for disease, to direct further testing, and to prognosticate. The relationship between diagnosis, prognosis and therapeutic response is complicated, as diagnostic categories may themselves be influenced by the ability to prognosticate and to direct therapy, but it is still true that medical testing serves a wide array of purposes in medical practice.

Medical tests can be valuable. They can be used to explain a patient's symptoms, predict who will develop symptoms, direct patients toward lifesaving therapy, and detect people who are infectious and pose a risk to others. Good tests, if used properly, can do wonderful things. However, test results can also lead to the wrong explanation of a patient's symptoms being given, to patients being told they are gravely ill when they are not, to being told they are not gravely ill when they are, to being denied potentially life-saving treatment, or to being given an unnecessary and potentially harmful one, to allowing infectious people to visit and infect vulnerable people, or to denying people access to their loved ones when they pose no threat to them. If used improperly, even good tests can do terrible things. This is why the performance of medical tests should be evaluated carefully.

The evaluation of medical tests is challenging. Studies evaluating tests are often deemed of poor quality, haphazard or incomplete.<sup>2</sup> “Poor quality of diagnostic studies is a recognized problem”.<sup>3</sup> Significant methodological challenges remain in medical test evaluation research, which has received less attention than methods for the evaluation of therapies. “Whereas the development of diagnostic technologies has greatly accelerated, the methodology of diagnostic research still lags behind that of evaluation of treatment”.<sup>4</sup> Philosophers have also recognized that the evaluation of diagnostic practices has received much less attention than the evaluation of therapeutic interventions. “Although much has been written on the role of randomized controlled trials and mechanistic reasoning in the evaluation of therapeutic treatments, philosophers of medicine have not yet turned their attention to the question of how diagnostic tests and procedures should be evaluated”.<sup>5</sup>

This lack of philosophical attention is regrettable. There are many questionable philosophical assumptions embedded in standard methods of diagnostic test evaluation, which have influenced and continue to influence research. These include assumptions about the characteristics of disease, and about the sort of thing diseases are. These are rather abstract considerations. Medicine is a practical subject, and people concerned with the practice of medicine may instinctively regard such abstract reflections as little use to medicine. Although understandable, I fear this attitude is foolish. Many things that people think exert a profound influence over what they do. This is true even for quite abstract considerations, which people may have never articulated or even thought about carefully. It is worth reflecting on this.

Consider a scene from the film “*In the Valley of Elah*”,<sup>6</sup> directed by Paul Haggis, in which two detectives are investigating a murder. A witness reports that they saw a green car parked near where the victim’s body was found on the night the murder took place. One detective, on this evidence, has instructed the police to search for this green car, as the owner might be the murderer. The other detective visits the crime scene and sees that this car was parked under a yellow streetlight. The second detective announces to the first that the police should not be looking for a green car, but rather for a blue car, because a blue car parked under a yellow light at night looks green. What has happened here is that the first detective, tacitly and perhaps unconsciously, accepted that the lighting conditions under which objects are viewed do not influence the perception of their colour. This is a rather abstract, theoretical belief, which the detective had probably never articulated or considered carefully. If asked, the detective may even deny that they hold such a belief. Nevertheless, the tacit acceptance of this belief had

---

<sup>2</sup> Verbakel et al. 2017

<sup>3</sup> Fontela et al. 2009

<sup>4</sup> Knottnerus, Buntinx, and van Weel 2008

<sup>5</sup> Kennedy 2016; see also Mebius, Kennedy, and Howick 2016

<sup>6</sup> *In the Valley of Elah* 2007

an important effect on the investigation, so much so that if it had not been articulated and addressed directly the police would have been looking for the wrong car. The real murderer may even be cleared of suspicion because their car is not green. Abstract, theoretical and unarticulated beliefs can influence practice in important ways, and this applies to medical practice as well.

Such beliefs can also embed themselves in practices, rather than in the minds of practitioners. The design of frameworks used to evaluate diagnostic tests might be influenced by rather abstract beliefs about the nature of disease held by the researchers who designed them. Subsequent generations of researchers may adopt these frameworks without themselves accepting the beliefs that inspired their structure. Even so, by using those frameworks, the beliefs may live on in the practice of test evaluation. Researchers may themselves hold several conflicting beliefs, which conflict with each other and with the beliefs embedded in the practices they adopt. Such conflicts produce the tensions in medical test evaluation that are the subject of this book.

This book is a philosophical reflection on the tension between understanding the characteristics of disease as homogeneous and unique, on the one hand, and understanding the characteristics of disease as heterogeneous and commonplace, on the other. This contrast has two senses.

The first has to do with how the characteristics of disease relate to each other. To understand the characteristics of disease as homogeneous is to believe that all medically interesting characteristics covary with disease status. So, knowing about disease status tells us everything we could want to know about the patient, from a medical perspective. In contrast to this, to understand disease as heterogeneous is to believe that medically interesting characteristics do not covary with disease status. So, knowing about disease status does not tell us everything we want to know about the patient, from a medical perspective. Borrowing language from Alvan Feinstein,<sup>7</sup> I describe the relationship between medically relevant characteristics in terms of a disease universe. A homogeneous disease universe is one in which all medically relevant characteristics covary, and a heterogeneous disease universe is one in which all medically relevant characteristics do not covary.

The second has to do with how the characteristics of disease relate to disease status. Medical tests measure a characteristic that is used to indicate disease status. To understand the characteristics of disease as homogeneous is to believe that all patients with the a disease status have the same tendency to have a characteristic used in a medical test. Accompanying this belief in the *homogeneity* of disease characteristics used in medical testing is a belief in their *uniqueness*, a belief that this tendency to have the characteristic is only found in patients with the disease. The belief that the characteristics of disease used in medical testing are homogeneous and unique has played

---

7 Feinstein 1967

an important role in medical test evaluation over the twentieth century. In contrast to this, since the new millennium there has been increasing recognition that patients with the same disease are much more *heterogeneous* than expected, and that characteristics used in medical testing are also *common* in patients with other diseases and in patients with no disease at all.

Accompanying this increasing recognition of heterogeneity has been an upsurge in calls made by medical researchers to improve how medical tests are evaluated. Traditionally, the evaluation of medical tests has focused on the evaluation of *diagnostic accuracy*: the ability of tests to distinguish people with a particular disease from people without it. In this book, the term ‘accuracy’ refers to this ability, unless stated otherwise. Today, medical test evaluation is still overwhelmingly focused on test accuracy studies.<sup>8</sup> Typically, this has been done by assessing a test’s *sensitivity* (the proportion of diseased patients with a positive result) and *specificity* (the proportion of non-diseased patients with a negative result). Researchers, however, have highlighted problems with traditional case/control designs for studies assessing diagnostic accuracy, which evaluate sensitivity and specificity in two separate populations, recommending that sensitivity and specificity be evaluated in the same population, because populations are unique and tests perform differently in different populations.<sup>9</sup> There have been growing calls to ensure that diagnostic tests are always evaluated in the clinical population in which they will be used.<sup>10</sup> Methodological reviews of how to conduct diagnostic accuracy studies started to emphasize the problem of patient heterogeneity, in an effort to improve the quality of such studies.<sup>11</sup> Several books have been published on how to evaluate diagnostic tests in the era of evidence-based medicine, as have papers on how to evaluate a medical test when disease status cannot be definitively established.<sup>12</sup> Researchers have called for more attention to be paid to evaluating the *utility* of medical tests, in addition to their *accuracy*, encouraging the use of randomized controlled trials to evaluate diagnostic tests.<sup>13</sup> Indeed, frameworks on how to evaluate diagnostic tests have recently been published that do not prioritize accuracy over utility.<sup>14</sup> Traditional test evaluation procedures have been subjected to increasing scrutiny in recent years and I connect this trend with the increasing recognition of patient heterogeneity.

Attempts to adapt traditionally accepted test evaluation procedures, to try to cope with the increasing recognition of how heterogeneous and commonplace these characteristics are, have produced significant tensions in test evaluation research. On the

---

<sup>8</sup> Verbakel et al. 2017

<sup>9</sup> Rutjes et al. 2005; Olsen et al. 2019

<sup>10</sup> Irwig et al. 2002; Irwig et al. 2008; Royal Statistical Society 2021

<sup>11</sup> Irwig et al. 2002; Bossuyt et al. 2003; Bossuyt et al. 2015; P. F. Whiting et al. 2003; P. F. Whiting et al. 2004; P. F. Whiting et al. 2011; Jaqueline Dinnes et al. 2005

<sup>12</sup> Knottnerus and Buntinx 2008; Newman and Kohn 2009; Reitsma et al. 2009

<sup>13</sup> Lord, Irwig, and Simes 2006; Lijmer, Leeflang, and Bossuyt 2009; Bossuyt et al. 2012

<sup>14</sup> Van den Bruel et al. 2007; Horvath et al. 2014

one hand, researchers treat the changes diseases produce as homogeneous and unique, whilst on the other hand wanting to recognize how heterogeneous and commonplace these changes are. Researchers evaluating medical tests are grappling with philosophical assumptions about the homogeneity and uniqueness of the characteristics of diseases, and this process deserves attention because these philosophical assumptions influence how medical tests are evaluated and affect medical practice.

My starting point is to highlight two strange and recurring contradictions in the medical test evaluation literature, which result from the tension between recognizing homogeneity or heterogeneity of disease characteristics. The first is that researchers claim a test can only be useful if it is accurate, whilst at the same time claiming that knowing a patient's disease status does not necessarily provide other useful information about the patient. The second is that some researchers believe sensitivity and specificity (indices widely used to evaluate test performance) are *fixed* properties of a test, which are applicable in many different clinical contexts, whereas others believe that sensitivity and specificity are *not* fixed properties of a test, which vary in different clinical contexts. These sets of claims are in tension with one another, and this tension is important. Accepting that the disease universe is homogeneous entails that a test can only be useful if it is accurate. If the disease universe is actually heterogeneous, useful tests may be rejected because they are not accurate, and worthless tests may be accepted because they are accurate. If the characteristics of disease are homogeneous and unique, then the sensitivity and specificity of a test will be fixed in all clinical contexts. If the characteristics of disease are actually heterogeneous and commonplace, then tests that perform well in one clinical context will be assumed to perform well in all clinical contexts even if they do not, and vice versa. Our understanding of what medical tests can be used to do depends upon our views about how patients are homogeneous and heterogeneous – about how patients are similar and different to each other. It is therefore important to reflect on how the homogeneity and heterogeneity of patients has been handled in medical test evaluation.

In chapter 1, I argue that believing that tests can only be useful if they are accurate, and that sensitivity and specificity are fixed properties of a test, reflect the underlying assumption that diseases produce homogeneous and unique effects in patients. Despite their widespread acceptance, neither of these beliefs is especially intuitive. The example of infectious disease shows that groups of patients who are infected, infectious, symptomatic, respond well to therapy, and who have a poor prognosis, are often not identical, meaning that a test that cannot accurately detect the group of diseased patients might be useful for detecting other patient groups. The toy example of a metal detector being used to find gold shows that believing sensitivity and specificity to be fixed properties of a test is far from intuitive, as the performance of a metal detector will obviously vary dramatically in different contexts. I describe how, in order to cope with patient heterogeneity, researchers have begun to understand disease as one of many “target conditions” that a test may be used to detect, and how they have start-

ed to investigate how medical tests perform in different clinical contexts.<sup>15</sup> The beliefs that knowledge of disease status is potentially uninformative about other “target conditions”, and that sensitivity and specificity are not fixed properties of a test, reflect the underlying belief that diseases produce heterogeneous and commonplace changes in patients. These ontological positions about the character of diseases are in tension, and this tension manifests in contemporary medical literature.

In chapter 2, I describe some medical research that further questions the assumptions of homogeneity and uniqueness of disease entities. These assumptions have been so widely accepted that it is worth documenting real life scenarios where they do not apply. Sensitivity and specificity have been found to vary for a wide variety of reasons, including differences in clinical population, the clinical pathway in which testing takes place, the timing of testing, the tests used to determine disease status, and the purpose of testing. Consequently, tests can be useful even if they do not detect disease accurately, and sensitivity and specificity are not fixed properties of tests. The assumptions that disease entities produce effects that are highly homogeneous and unique do not bear scrutiny.

Given this, why have researchers traditionally assumed that diseases are highly homogeneous and unique? In chapter 3, I connect these assumptions with the notion of specific disease entities, defined according to their common aetiology. The first test for which sensitivity and specificity were assessed was the Wassermann reaction, a serological test for syphilis, in the early twentieth century.<sup>16</sup> The notion of specific disease entities was popular in this period, and it shaped the development of the Wassermann reaction. The Wassermann reaction set a precedent for the development of other diagnostic tests, popularizing test evaluation practices that incorporated notions of specific disease entities and embedding assumptions of homogeneity and uniqueness into test evaluation practices. I show that physicians thinking about how to diagnose cancer in the mid-twentieth century were influenced by the perceived success of the Wassermann reaction and believed that diagnostic tests would be developed in cancer when the discovery of specific aetiologies for cancer could produce sufficiently homogeneous disease categories. Following this, clinical epidemiologists began trying to use computer-aided diagnostic procedures and developed tables to display the prevalence of certain disease characteristics in each disease, preserving the notion that each disease is highly homogeneous, even if the interest in aetiology was diminished. Researchers in the last quarter of the twentieth century recognized that sensitivity and specificity of tests did appear to vary, but explained this variation as the result of the imperfect allocation of disease status, rather than as the result of patient heterogeneity. In the early twenty-first century, more and more researchers began reacting differently

---

<sup>15</sup> Bossuyt et al. 2003

<sup>16</sup> Binney, Hyde, and Bossuyt 2021

to variations in test performance, explaining them in terms of patient heterogeneity, and calling for reform of traditional ways of evaluating medical tests.

Although researchers throughout the twentieth century often recognized that the changes produced by diseases were heterogeneous and commonplace, I identify several instances where researchers still chose to treat diseases as if they were homogeneous and unique and discuss why they did so. I show that assumptions about how homogeneous and unique these characteristics are have had a profound influence on test evaluation throughout the twentieth century. This is not to say that these were poor assumptions to make. Assumptions about specific disease entities with homogeneous and unique characteristics have been at the foundation of tremendous medical achievements that have been made in the twentieth century. Even so, that these assumptions have been beneficial does not mean that they cannot also be damaging.

The goal of this history is not to be comprehensive. The development of medical testing cannot be represented as one single conversation in which all interested parties were participating. Medicine is far too disparate a discipline for this to have happened. The episodes I highlight have been chosen to illustrate how questions of homogeneity and heterogeneity have been handled by some influential actors over the twentieth century, so that comparisons can be made between them. These comparisons highlight that different actors took different things for granted and had their own different reasons for doing so. Even so, these episodes are not isolated from one another, as the actors in earlier episodes inform the beliefs and practices of actors in later episodes to some extent. History can both facilitate an awareness of the things we today take for granted and help explain why we do so.

In chapter 4, I discuss tensions that exist in traditional medical frameworks for medical test evaluation. These traditional frameworks are hierarchical, requiring that a test demonstrate its value at lower levels in the hierarchy before it is evaluated at higher levels in the hierarchy. Distinguishing obvious or severe disease from health is deemed less difficult than distinguishing mild or occult disease from health, or from other diseases. Hence, tests that cannot distinguish obviously diseased patients from well patients in selected populations are assumed to be unable to distinguish patients with mild or occult disease from healthy people, or from people with other diseases. Similarly, if a test cannot distinguish people with a disease from people without a disease, then it is assumed that it cannot be useful. I argue that all these hierarchical practices rest on the intuition that disease characteristics are highly homogeneous. And yet, the insistence that test performance needs to be evaluated in different populations implies that the patients in these different populations are heterogeneous. I also argue that insisting that the utility of a test needs to be evaluated in addition to its accuracy and the efficacy of treatment implies that patients with the same disease status are heterogeneous. So, there is a tension between homogeneity and heterogeneity in these hierarchical frameworks. Using several examples drawn from medical practice, I expose and explore these tensions, arguing that if the heterogeneity of patients with

the same disease is to be taken into account, hierarchical frameworks may need to be abandoned.

Finally, in chapter 5, I describe some recent examples of how assumptions of homogeneity and uniqueness of diseases cause problems in medical practice. Diagnostic tests for tuberculosis have been evaluated in contexts where the performance of these tests appears excellent, and yet in practice their performance has been very poor. This has led the World Health Organization to offer its first negative advice, that these tests should not be used, and to the banning of these tests by the Indian government. During the recent COVID-19 pandemic, Public Health England made an evaluation of an antigenic test for the disease, finding that it had a high sensitivity, when in practice it had a low sensitivity. These examples show that how to handle the tension between understanding the effects of diseases as homogeneous and unique, or heterogeneous and commonplace, is still a live issue today. These tensions are evident even in Cochrane Collaboration systematic reviews of tests for COVID-19 disease, which represent the highest quality research into these medical tests. Even in these reviews, researchers acknowledge the importance of exploring patient heterogeneity, whilst often ignoring it.

In focusing on tensions between homogeneity and heterogeneity in medical test evaluation, this book is unlike any other in the field. There are many texts in clinical epidemiology that describe how to carry out a diagnostic accuracy study, conduct the relevant statistical analyses (such as plotting receiver operating characteristic curves and multivariable regression analyses), watch out for biases, and conduct research into utility and cost-effectiveness. None of these texts, however, grapple with how basic ontological assumptions concerning homogeneity or heterogeneity of disease characteristics influence the evaluation and interpretation of indices such as sensitivity and specificity. This book has the advantage of taking an historical approach to these questions (in chapter 3). Attending to the development of how medical test evaluation was thought about and practised allows us to see that different assumptions about disease have been adopted by different people at different times. The comparison of these different approaches highlights that we too in the present are making assumptions about the nature of the diseases we are trying to detect and invites us to reflect upon them. This is an instance of integrated history and philosophy in a complementary mode, as championed by Hasok Chang,<sup>17</sup> and draws upon the comparative epistemology championed by Ludwik Fleck.<sup>18</sup>

As pointed out by Ashley Graham Kennedy in her welcome book on diagnosis,<sup>19</sup> the philosophy of diagnosis has received very little attention. There are exceptions to

---

<sup>17</sup> Chang 2004; Chang 2012

<sup>18</sup> Fleck 1979

<sup>19</sup> Kennedy 2021

this,<sup>20</sup> but this important aspect of medical practice has truly been neglected by philosophers. How to evaluate and use medical tests has received even less attention. An exception to this is chapter 2 of Kennedy's book. However, the tensions I am addressing here are also present in Kennedy's book. On the one hand, Kennedy largely takes the homogeneity of disease characteristics for granted, saying that "in medical practice, accurate diagnoses are always the first goal because without them, both treatment plans and prognostic predictions are bound to fail".<sup>21,22</sup> Here, the disease characteristics are so homogeneous that prognostic and therapeutic information can only flow from knowledge of an accurate diagnosis. However, Kennedy also recognizes that an accurate diagnosis does not necessarily provide prognostic or therapeutic information,<sup>23</sup> thereby treating disease characteristics as heterogeneous. Kennedy does not discuss whether sensitivity and specificity are fixed properties of a test, but at least some of the tensions I discuss here are present in her work. This is not surprising, as these tensions exist in the medical literature, in test evaluation practices, and in the minds of medics, and Kennedy has paid close attention to medical practice. Furthermore, Kennedy's book is not designed to be a comprehensive philosophical account of diagnosis, as it has much too wide a scope for that. This book has a much narrower scope, being focused on medical test evaluation alone. Even so, this book is not designed to be comprehensive either. It is designed to take the invitation made by Kennedy<sup>24</sup> for others to join her and delve more deeply into an area of medical diagnosis. I hope that this discussion of the tension between homogeneity and heterogeneity in medical test evaluation will be useful and productive to medical and philosophical practice.

Of course, there is a sense in which any individual patient is unique, and unlike any other patient. From this point of view, patients are heterogeneous, and all assumptions of homogeneity are artificial distortions that ignore differences between patients. However, if medicine does not treat some patients as the same as each other, and as homogeneous, in some sense, then medicine cannot learn from past experience about how to treat patients in the future. Medicine can be understood as a search for useful ways of treating heterogeneous people as homogeneous in order to help them. Very roughly, throughout the twentieth century patients with the same disease status have

---

<sup>20</sup> K. Sadegh-Zadeh 2000; Kazem Sadegh-Zadeh 2011; Kazem Sadegh-Zadeh 2015; Stanley and Campos 2013; Lalumera and Fanti 2021

<sup>21</sup> Kennedy 2021, 23

<sup>22</sup> Kennedy also says that "Unless a diagnostic test is considered to be accurate, no physician or patient will be (or should be) interested in the results. So we can readily state, then, that accuracy is the minimum bar that all diagnostic tests must reach in order to merit further consideration" (Kennedy 2021, 25), and "As discussed previously, it is generally agreed upon that unless a test is at least relatively accurate (no diagnostic test is 100 % accurate), it will not be useful in clinical practice, and there is thus no point in performing it" (Kennedy 2021, 26). These claims reveal a deep commitment to the homogeneity of the characteristics of disease.

<sup>23</sup> Kennedy 2021, 31–32

<sup>24</sup> Kennedy 2021, viii

been treated as homogeneous, as the same as each other for all intents and purposes, and different to everyone else. Whilst this has been a powerful and productive assumption, it is causing problems in medical test evaluation. How patient homogeneity, heterogeneity, uniqueness and commonality have been handled in medical test evaluation deserves much more philosophical attention.