Chapter 9
Scope Validity in Medicine

Lara Keuck

9.1 Introduction

If a test measures what it means to measure, it is deemed "valid." First defined in this way in psychological research (Kelley 1927), the concept of validity has pursued a steep career. Since at least the mid-twentieth century, the ideal of validity has been theorized, debated, translated into methods, and used to regulate and (de-)legitimate knowledge concerning health and disease. For instance, a specific rodent model of chronic mild stress was considered one of the best validated animal models of depressive disorder in humans according to existing concepts of validity (e.g. Willner and Mitchell 2002). However, clinical trials on therapeutics that had been successfully tested in the animal model failed. The reason for this failure in the human context has been attributed to the fact that only a small portion (and therefore a financially uninteresting market) of patients who are diagnosed with depression suffer from a subtype of the disorder for which this model is a good predictor (Belzung 2013). The clinical trial population was not stratified in a way that allowed to test whether or not the drug works. Put differently, the experimental design of the preclinical model restricted the successive domain of application of the research results. This case of translational failure can be analyzed in several ways: we can question the meaning of 'best validated model' if the animal model cannot be adequately extrapolated to clinical trials on depression. Or we can blame the pharmaceutical company's marketing-oriented selection of too broad inclusion

L. Keuck (✉)
History and Philosophy of Medicine; Faculty of History, Philosophy, and Theology, and Medical School OWL, Bielefeld University, Bielefeld, Germany

Max Planck Research Group "Practices of Validation in the Biomedical Sciences", Max Planck Institute for the History of Science, Berlin, Germany
e-mail: lara.keuck@uni-bielefeld.de

M. Schermer, N. Binney (eds.), *A Pragmatic Approach to Conceptualization of Health and Disease*, Philosophy and Medicine 151,
https://doi.org/10.1007/978-3-031-62241-0_9

criteria for undermining the model's validity. Both approaches are fair enough, yet the blame game that often results can easily overshadow that validity is never unmediated, never absolute. Mismatching of scopes are not just (though also) a problem of polemics and the rhetoric of big pharma or overpromising biomedical research. We lack an understanding of the scientific activities involved in capturing and evaluating how well the scope of an experiment—the *actual domain of application* of the results of preclinical research—fits to its *intended target domain of application* in the clinical context.

This chapter analyzes mismatching disease operationalizations as challenges to validity in biomedicine, and introduces the new concept of scope validity to capture this problem. It combines an adequacy-for-purpose view towards validity (e.g. Alexandrova and Hybron 2016; Feest 2019; Parker 2020) with a pragmatist and particularistic perspective on disease concepts (e.g. Demazeux and Keuck 2023; see also Binney et al., Chap. 2, in this volume for a pragmatist perspective on disease concepts; Kusch, Chap. 5, in this volume for differentiating pragmatism from relativism; Binney, Chap. 7, in this volume for conceptualizing change in disease operationalizations). The chapter proceeds as follows: the second section focuses on mismatching disease operationalizations as a missing link in the evaluation of animal models of human mental disorders. Against this background, I clarify how my notion of scope validity differs from existing concepts of validity, in particular construct validity, external validity, and predictive validity. In the third section, I advocate much in the spirit of practical concepts of disease for a relational epistemology to biomedical objects of inquiry. I argue for relational epistemology as a philosophical framework for capturing the extent to which (and the conditions under which) the relata of a specific animal model, a clinical trial design, and the diagnosis in clinical guidelines match. This line of argument builds on my particularistic perspective, which side-steps all-encompassing validity theories and general philosophical theories of disease, while being attentive to the diversity of validity and disease theories that are at work in every single study design. Against this background, I argue for the potential of a philosophy of science in practice approach to identify existing medical scientific methods that could be analyzed as responding to problems of scope validity. For instance, some forms of retrospective epidemiological studies and reverse translation trials in animal models (testing effective clinical interventions in animals) might be understood as instances of 'scoping methods,' which provide us with information on the (mis-)matching of disease operationalizations in different research and application contexts. In the concluding section, I address the functions we might ascribe to scope validity: as a tool for evaluating study designs in translational medicine, as a description of how knowledge generation within one biomedical context conditions the way in which a medical problem needs to be identified in another context, and as an analytic category for studying scientific methods of matching scopes across research contexts. I conclude with a common thread between the philosophical questions that scope validity raises: the adequacy of approaches to medical research.

9.2 Validity, Scope, and Scope Validity

This section introduces scope validity. I first analyze the role of abstract targets (or constructs) in validity concepts (Sect. 9.2.1). I will then examine the limitations of this approach for evaluating animal models of human diseases (Sect. 9.2.2). Against this background, I discuss the representational scope of models in biomedical research and present scope validity as a complementary conceptual tool to identify the target population to which a research result might be best generalizable (Sect. 9.2.3).

9.2.1 Validity Concepts and the Guiding Ideal of a Construct

Validity has been debated for almost a century, especially in the psychological sciences. Most validity theorists take the educational psychologist Truman Lee Kelley's 1927 dictum as point of departure: "The question of validity would not be raised so long as one man uses a test or examination of his own devising for his private purposes, but the purposes for which schoolmasters have used tests have been too intimately connected with the weal of their pupils to permit the validity of a test to go unchallenged (. . .) *The problem of validity is that of whether a test really measures what it purports to measure*" (Kelley 1927: 14, my italics). Validity seems to involve "the acceptance of a set of operations as an adequate definition of whatever is to be measured" (Bechtoldt 1951, 1265, quoted in Cronbach and Meehl 1955, 282). Or at least this is the case for a specific understanding of validity. Indeed, this was the worry of Lee Cronbach and Paul Meehl, the heads of the Committee of the American Psychological Association that was tasked to formulate *Technical Recommendations for Psychological Tests*. They suggested an elaborate terminology of different kinds of validity, naming their "chief innovation" the introduction of a new term that they called "construct validity": "Construct validity is not to be identified solely by particular investigative procedures, but by the orientation of the investigator. (. . .) When an investigator believes that no criterion available to him is fully valid, he perforce becomes interested in construct validity because this is the only way to avoid the 'infinite frustration' of relating every criterion to some more ultimate standard" (Cronbach and Meehl 1955, 282). They suggested a new concept, namely that of construct validity, to give "investigators" a possibility to address a specific kind of doubt: not a doubt about the performance of a test, but about its informativeness about an abstract target.

The concept of construct validity becomes more intelligible when taking into account the nature of 'constructs.' Ken Schaffner (Forthcoming: 1) defines constructs as concepts that "refer to entities that are general, abstract, and putatively explanatory. Examples include notions such as intelligence, working memory, gamma frequency oscillation circuits, normal and abnormal personality types, disorders such as schizophrenia, and even the 'self.'" If a test has a high construct

validity, it is highly informative about the abstract entity in question. A valid test can be understood as providing evidence for the reality of the construct (if we can measure intelligence, it exists), and/or as being a good way to test the manifestation (e.g. of intelligence) in an individual that allows for drawing conclusions that are also of relevance outside of the test context.

Psychometricians, who were the first to introduce and broadly apply notions of *construct validity*, for instance with regards to psychological testing of personality traits or intelligence, have developed a nuanced terminology. Keith Markus and Denny Borsboom (2013: 3) define a *construct* as a "property tested or intended for testing," which "assumes a substantive interpretation of this property." The semantic representation of this property is then the *construct label*. Since the "researchers do not directly observe" the property, the psychometricians treat it as a *latent variable*, which allows them "to represent statistical relationships with some latent variable, whatever it may be, without specifying the substance of that variable."

According to Schaffner (2012, Forthcoming), the introduction and use of validity concepts in psychometric, psychiatric, and animal model research contexts have given rise to quite different discussions with varying underlying philosophical commitments to laws, pragmatism, and reductionism. However, Schaffner also stresses that the notion of a construct as an abstract entity has been central to all three of these contexts, even if, for instance, Robins and Guze's (1970) criteria on how to assess whether a diagnosis of schizophrenia was valid did not at all refer to Cronbach and Meehl's term of construct validity. Moreover, he seems to agree with Cook and Campbell (1976) who "asserted that C[onstruct] V[alidity] was involved *whenever* one dealt with causes and outcomes." (Schaffner Forthcoming: 2). It is a fair assumption that construct validity served a regulatory function for the many other validity concepts — internal, external, predictive, descriptive, aetiological, face, etc. (Sect. 9.2.2) — that had been introduced and discussed in the last 65 years. At stake was the question of how well a certain model or test hit the abstract target of inquiry, be it with respect to representing its pathophysiology, determining its relationship to a latent variable, or to developing a screening device for drug testing. While the plurality of validity concepts indicate an awareness of the various aims and interests in assessing the hitting of the target disease entity, the practical definition of the target mark, and the fitting of different definitions within contexts of experimentation and contexts of application remained undertheorized. Discussions on the validity of animal models for human diseases illustrate exemplarily why this concern matters.

9.2.2 *The Logic of Validation in Animal Models of Human Diseases*

Within the field of animal-based modelling of human diseases, most researchers have used variants of three suggested validity concepts: predictive validity, face validity, and construct validity. Yet, there is no homogenous use of these concepts

and their derivatives—not even in a comparably confined field, such as that of animal models of human depressive disorders (see Belzung and Lemoine 2011). In general, a high *predictive validity* denotes a high "human-animal correlation of therapeutic outcomes," that is to say, pharmacological (or other interventionist) therapeutic effects in humans can be reproduced in the animal and vice versa (ibid.: 3). A high *face validity* of an animal model means that it exhibits a "phenomenological identity" to the human disorder, which is mostly understood in terms of "an attempt to mimic diagnostic criteria of the psychiatric conditions, such as those listed in the American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders." (ibid.: 4). A high *construct validity* means that the animal model is informative about the human disease in the sense that the model can be used to gain knowledge about the disease entity in question.

But how do we know what qualifies as a legitimate instance of this construct? Models often fulfil a seemingly paradoxical role, the trained psychologist turned animal researcher Paul Willner noted while revising the face validity criteria of animal models of depression that were first proposed by McKinney and Bunney (1969). Along with his revisions, Willner introduced new measures of predictive and construct validity (Willner 1984), which he argued were necessary updates because, for one, "in relation to animal models of depression, similarity of aetiology and biochemistry are unsuitable as validating criteria since they are themselves the subject of intense research and speculation." (ibid.: 1). For another, Willner's update was motivated by his perception that scientific progress in depression research had led to new hypotheses regarding the interrelation between environmental factors and endogenous depression as well as more elaborate experimental set-ups to induce and test behaviors, for instance, the animals' reactions to 'chronic stress.'

This example, and, more generally, the plethora of validity concepts that scientists, psychiatrists, and philosophers have elaborated in the last decades reflect the manifold interdependences between determining the explanatory role, the predictive power, and the representational scope of a given test or model (see, e.g., Kendler and Parnas 2012). Perhaps best known and most discussed in philosophy of science is the differentiation between internal and external validity (e.g., Cook and Campbell 1979): research results are *internally valid*, when they are reproducible and significant within the confined parameters of a controlled test, but need to prove their *external validity* outside of the controlled, experimental setting in real-world contexts (see also Guala 2003; Cartwright 2009). Extrapolation and external validity have been the subject of many philosophical inquiries into application-oriented sciences, some of which have motivated normative conclusions on how science should work in order to be useful for society (e.g., Kitcher 2003; Cartwright 2009). Yet, the chronological and epistemic order that presumes that internal validity always precedes external validity is challenged in biomedical research, which operates in a more iterative mode (Huber and Keuck 2013). Biomedical research does not start at the bench and end in the clinics; the material and conceptual transfers are multidirectional.

This iterative go-between of clinical and laboratory demands and insights is particularly evident within animal models of diseases that are thought to occur

only in humans, such as Alzheimer's disease, a neurodegenerative disease leading to dementia and death. The establishment of a mouse model begins with a reverse translation from bedside to bench, often including transfer of genetic material from human patients to laboratory animals. It necessitates the selection of clinical symptoms (e.g., memory deficits, but not personality changes) and their translation into test procedures for animals (e.g., behavioral testing of mice's memory deficits in the Morris Water Maze). After establishing and characterizing the animal model, pharmaceuticals are tested in these in-bred animals. The conclusions in the lab legitimate whether the drug should be tested in clinical trials on humans.

The zigzagged logic of animal modelling has implications for thinking about what it means to 'hit' the target of inquiry. The resort to an abstract disease construct has clouded rather than facilitated the assessment of the representational scope of an animal model with regards to human patients. Alzheimer's disease is perhaps a particularly strong case in point with its unknown aetiology and its ambiguous definition (Huber and Keuck 2013; Keuck 2020; Daly and Keuck 2024). The first mouse model that exhibited both a (nowadays debated) histopathologic hallmark of the disease (amyloid beta plaques) and memory deficits (Hsiao et al. 1996) had been established through the transfer of genetic material of the so-called Swedish mutation. This genetic mutation had been characterized within a human genetic field study that had traced families in which severe, early onset forms of dementia had occurred throughout generations. The geneticists that had isolated (and later patented) the Swedish mutation acknowledged that Alzheimer's disease was "genetically heterogeneous" (Mullan et al. 1992: 345). However, the mouse model was not presented and evaluated as a model that might provide more insights into the devastating illness of this Swedish family, but as a model for Alzheimer's disease in general. In the past 25 years, several hundred further mouse models for Alzheimer's disease have been established and elaborately validated, but in terms of translational research this approach did not provide for successful extrapolations. Just as in depression research, Alzheimer's researchers working with mouse models have blamed the clinical trial designers for redefining the medical target: "The nosology of A[lzheimer's] D[isease] keeps shifting, the consequence of not knowing its etiology. This situation makes it difficult to place mouse models precisely into human context and demands an adaptive framework for utilizing mice as models of the human disease." (Ashe and Zahs 2010). In other words, the clinical redefinitions have made the animal researchers' validation work of Alzheimer mouse models invalid.

The zigzagged logic of validating animal models for human diseases may remind us of the philosophical characterization of so-called looping effects. Originally, Ian Hacking (e.g., 2007) has described looping effects that stem from classified people's reactions to the way they have been classified, which can result in a change of this classification (e.g., of autism or of homosexuality in psychiatric manuals). Such 'moving targets' could be seen as one cause for a subset of classificatory shifts. The problem that I address in this paper, however, encompasses many more kinds of mismatches between an implicit or explicit definition of a target of inquiry in one setting (e.g., a particular lab) and in another (e.g., in a clinical trial, or in a general

physician's practice). Jackie Sullivan (2009) has argued in a similar vein that it might turn out to be difficult to assess what neuroscientific studies can tell us about memory in general, when the protocols that are used in different laboratories to operationalize memory differ so strongly from each other that it is no longer clear whether they actually relate to the same phenomenon.

With regards to Alzheimer's disease, one strategy — that is currently propagated by the National Institute of Aging — to solve this problem of the shifting target is to bind the construct label Alzheimer's disease to a measurable variable like the occurrence of amyloid plaques in the brain (Jack et al. 2018). However, this strategy has several problems: as it reduces the mental illness to a biomarker, it is likely to be overinclusive with respect to false positives, because heightened values of amyloid beta also occur in people who never develop clinical symptoms. It also deprivileges alternative aetiological hypotheses, which might, according to some epidemiologists, account for a significant proportion of cases that are contemporarily diagnosed as Alzheimer's disease (e.g., Glymour et al. 2018). The overarching problem is that neither contemporary epidemiological nor biomarker approaches provide sufficient grounds for defining Alzheimer's disease unequivovally: similar to what Paul Willner described with respect to the challenges of modelling depression in mice when we do not really know what qualifies depression in humans, we are faced also in the case of Alzheimer's disease with an epistemological underdetermination of the target of inquiry (Daly and Keuck 2024). What does it mean in such cases to deem a model valid?

9.2.3 Scope Validity

My suggestion is to take a step back from the definitory muddle (or warfare, in some cases) that surround many abstract constructs, and think about a measure that better qualifies the actual scope of a given test or model. With respect to biology and biomedicine, most scholars have identified the representational scope of a model with the degree to which a model and its target share essential properties or functional processes and therefore are instances of the same 'general biology' (see e.g., Burian 1993; Schaffner 1998; Keller 2000; Ankeny and Leonelli 2011; see also Steel 2008 for a defense of 'comparative process tracing' to grant successful extrapolation even if properties between the model and target differ). However, extrapolation and representation might, at least in some cases, work significantly differently within models for general biology as compared to biomedicine: in biomedicine, the relationship between experiment and application is one of substitution (i.e. animals replace human patients) rather than, necessarily, an exemplification of general biology (i.e. animals represent general patterns of interest; see Rheinberger 2006a; Huber and Keuck 2013; Germain 2014; Green 2024). For example, when xenografts or human genetic material are used to generate humanized animal models (as is the case in many Alzheimer mouse models), the question is not only one of how conclusions drawn from an animal disease can be extrapolated

to sick humans, but also which aspects of the human disease can be instantiated in the animal not least since Alzheimer's disease is thought to not naturally occur in mice.

In biomedicine, the target of the representational scope must be qualified not just regarding the comparability (be it the similarity or the possibility for comparative process tracing) of animal and human physiology, but also with respect to two further dimensions. First, we need to consider the degree to which this model can account for relevant aspects of human illness, recovery, or even the side effects of pharmaceuticals. For instance, weight gain as a side effect of a person taking antipsychotic drugs might be observable in animals but not the development of depressive symptoms due to the experience of the social stigma of obesity. To assess the psychological harmfulness, aspects of social (human) life need to be taken into consideration that are abstracted away in most experimental settings. This dimension relates to a model's face validity, i.e., the phenomenological similarity between model and target, whereby here face validity includes so-called patient-relevant outcomes. The second additional dimension that needs to be considered when determining the representational scope of a model in biomedicine is connected to the "reference class problem" (see, e.g. Hájek 2007): the Alzheimer's mouse mentioned above was potentially a much better model for the disease running in the Swedish family from which the mutated genetic material was transferred than it was for all people diagnosed with Alzheimer's disease in the 1990s. Similarly, the reference class of depression that the chronic mild stress rodent model is best compared to does not comprise all incidences of major depressive disorder, but a subclass that consists of people who developed depressive symptoms associated with stress. This dimension has some commonalities with variants of aetiological validity. However, it does not necessarily need to be based on a causal hypothesis of disease manifestation, which sometimes, but not always motivates stratification practices of delineating diagnostic groups. In psychiatry, so-called transnosographic and theranostic approaches have been adopted to suggest some re-classifications of mental illnesses (for an example, see Guessoum et al. 2020; for a critical assessment of a "precision psychiatry" approach, see Tabb and Lemoine 2021). Besides psychiatry, such re-grouping practices are much discussed and used in oncology, where, for example, umbrella or basket trials of cancer treatments cut across the traditional organ-specific classifications of neoplasms, and group traditional diagnoses together in new ways — with new chances and challenges for study trial designs and their implementation (e.g., Strzebonska and Waligora 2019). One of the challenges is akin to the missing link between assessments of the validity of animal models and successful translations into general health care practice that I elucidate in this paper: we need a measure of adequacy to ascertain how good the fit is between a given model or study population and the diagnoses in the "real world" patient population. This means to move from the abstract idea of a disease entity to concrete practices of identifying diseases in a given local context — and moving from a model's or trial design's construct validity to their scope validity in relation to a concrete context of application.

The idea of validity as a relational concept is not new, but to my knowledge it has not yet been elaborated in the medical sciences without assuming that there is an abstract disease entity that can be better or worse hit. For instance, Paul Willner once defined construct validity as "a theoretical account of the disordered behavior in the model, a theoretical account of the disorder itself, and a means to bring the two theories into alignment" (Willner 1994, quoted in Belzung and Lemoine 2011: 5). Other researchers even "mentioned the similarity of etiology, but also an interesting criterion that was unfortunately abandoned: the precision of the sub-nosographic entity ('Does the laboratory model describe (. . .) a naturally occurring psychopathology or only a subgroup?')" (Belzung and Lemoine 2011: 3, quoting Abramson and Seligman 1977). In the past two decades with the advance of -omics, big data analysis, and dimensional approaches to psychiatric classification, most prominently the Research Domain Criteria (RDoC) of the National Institute of Mental Health, the focus of many researchers has shifted away from abstract disease entities and towards modelling more fine-grained in-group differentiations, for instance between 'good' and 'poor responders' to antidepressants in mouse models (e.g. Herzog et al. 2018; for RDoC as a challenge to the predominance of diagnostic kinds in psychiatric theory, see Tabb 2017; Solomon 2022; Demazeux and Keuck 2023). Thus, there has been an increasing interest in refining a model's representational scope in practice, but this has not been theorized vis-à-vis the scientific validity concepts that still guide the choice of and warrant the extrapolation from animal models.

Concepts of validity that resort to an abstract entity cloud the fact that we lack a conceptual tool to capture mismatches between the scope of an animal model and the scope of a clinical trial. Catherine Belzung's (2013) frustrating conclusion of a failed clinical trial that had not been able to reproduce the effects of a pharmaceutical in human patients, which had been preclinically tested in her best-validated animal model of a specific form of depression, could be re-read as a call for taking the implications more seriously that the representational scope of the model has for defining appropriate inclusion and exclusion criteria in the clinical trial design.

Let me close this section with a working definition of scope validity:

(SV) *Scope validity denotes the matching between the target as operationalized in the setting of experimentation and the target as operationalized in the setting of application.*

In this understanding, the scope validity of an animal model in relation to a clinical study would include an assessment of the conditions of the particular clinical trial, which are best fitted to allow for testing the preclinically tested drug's mode of action in the human context. Importantly, scope validity is not identical to external validity, but rather an additional tool to *refine the frame under which external validity is assessed*.

In other words, we could describe a failure due to problems of scope validity as occurring when the domain of successful applicability is different from the domain of application that the intervention was tested on. Consider this definition from biomedical researchers Bert 't Hart et al. (2018) who apply validity terms to animal models of Multiple Sclerosis (MS): "External validity: represents the extent to which

the observed effect of a treatment in an animal model can be generalized to the *total* MS patient population." (ibid.: 263, my italics). In contrast to this total generalizability of external validity, scope validity would then represent the matching between *MS as it is diagnosed* in the *studied* patient population and *MS as it is modelled* in the animal. Importantly, I do not assume one specific form of relevant similarity between a target as it is modelled and a target as it is diagnosed, when I refer to the matching of targets. The question of what makes a good match rather is a central question that the focus on scope validity helps to address (see the next sections). In general, if the scope validity is high, this should imply that the translational set-up is, given the current scientific knowledge, well suited to test whether the preclinical results could be extrapolated to the human patient study group. Scope validity does in and of itself not capture the total generalizability of the intervention's effect, but rather the matching of a model and a specific clinical trial design. This has important consequences, because if we take 't Hart and colleagues' definition of external validity as stand-alone measure, we would need to disqualify animal models that only allow for generalizability for a small set of the total patient population as having a weak external validity. Yet, as has been argued in the case of clinical trials on antidepressants, there might be good candidates that could be effective for a subset of the patient population. If we took this group as reference class for evaluating external validity, the external validity would be presumably much higher than if we took the group of all people who receive a diagnosis of depression. At the same time, this need not mean that we should change the diagnostic criteria of a given disease altogether. In contexts beyond drug testing, for instance the assessment of socioeconomic factors that impact (mental) health, it might be more adequate to sample the patient population in a different way. As I have argued elsewhere in more detail, there are good reasons for taxonomic and explanatory pluralism in medicine, but it demands additional measures that check what transferred data or translated results exactly refer to (Kutschenko 2011a, b). Scope validity responds to this task.

9.3 Towards a Relational Epistemology

In this section, I contextualize my approach to scope validity within a relational epistemology that is based on a particularistic perspective on disease (Sect. 9.3.1). I argue that this approach allows to ask new philosophical questions about specific scientific methods that respond to the challenge of scope validity (Sect. 9.3.2).

9.3.1 A Particularistic Perspective on Disease

My definition of scope validity is grounded on a relational approach to the objects of medical research: this means to not compare how well a given practice hits an ideal

disease entity, but how well the target of one practice fits the target of another practice for which it attempts to provide a solution. This implies, for example, to not take for granted that practices of diagnosing disease or pursuing clinical trials carve nature at similar joints. The goal of a relational approach is to critically assess the extent and ways in which, for instance, the disease target that is operationalized when testing a drug in the highly-controlled setting of a clinical trial fits to the disease target that is reflected in the diagnostic practices of a primary health care setting (i.e. the context of application for which the context of testing attempts to provide a solution). In the trial, patient groups are often selected with expensive diagnostic technologies, such as PET-neuroimaging, but these technologies of identification are not available, affordable, and possibly desirable in all health care settings in which the experimental knowledge shall be put to practice. This is in those cases problematic, in which the ways of carving out the target in the experimental setting leads to a meaningfully different patient group composition than the one to which the knowledge is translated. In these cases, the practices identify different types though they are said to refer to the same disease. A relational theory acknowledges that in every single context, in every single laboratory, on every single occasion, a concrete manifestation (in an individual patient, in a clinical population, or in a model system of biomedical experiment) must be newly attributed to an abstract phenomenon-of-interest (like 'disease x' or 'memory', see Sullivan 2009; Feest 2011; Meunier 2012; Hauswald and Keuck 2017; Huber and Keuck 2017 for examples that are, however, not analyzed with explicit reference to a relational theory). Much in line with Gaston Bachelard's concept of a *phénomènotechnique*, the technologies and experimental procedures that are applied to make the phenomenon-of-interest examinable within the confines of a given research context impact the delineation of the target object (see Rheinberger 2006b). The very particular target objects of biomedical experiments therefore do not precede research although they aim to answer a question that is raised by a reality that exists outside of the laboratory.

The degree to which this particularistic perspective matters for successful translation of medical research results across local settings is a case-by-case empirical question. Some positively tested interventions into medical issues may require thorough knowledge and strict adherence to the precise rules of operationalization applied to the study. In other cases, there might be more tolerance.

The general approach of turning philosophical attention to practices of research is in line with a methodological development to characterize the generation, translation, and assessment of scientific knowledge in the real (read: social, complex, messy) world (see e.g., Wagenknecht et al. 2015). A main strand of research within this field of study has been the examination of how value judgements and divergent interests define the aim of a given research enterprise and thereby affect the design and evaluation of scientific studies (e.g., Longino 2002; Carrier 2004; Douglas 2009; Solomon 2015). "Identifying these features of a local epistemology, particularly the assumptions and values that link methods to kinds of knowledge sought, is a matter not just of picking out the methods and standards that link data to hypotheses in research articles but of reconstructing them from an analysis of the context of

inquiry: correspondence; accounts of controversy and of interventions in controversy; study of institutional settings, priorities, and constraints." (Longino 2002: 187).

The turn to local epistemologies of medical research has given rise to the acknowledgement that the multitude of sub-disciplines (e.g., anatomy, epidemiology, pharmacy) within medicine as well as the scientific approaches to medicine make use of various epistemological frameworks and metaphysical assumptions regarding theories of disease(s) (see e.g., Lemoine (2011) for an elaboration on the general claim with respect to explanations of disease). Anya Plutynski (2018) recently inquired with respect to the conclusions of Marta Bertolaso's (2016) study on the multitude of understandings of cancer, whether "we should consider giving up the very idea of general theories of cancer." It is not clear if the different theories relate to the same object. Preclinical studies might give rise to disease ontologies in the plural — just as Annemarie Mol has argued with respect to arthrosis within medical practice (Mol 2002). If this is the case, the much-discussed epistemological question of whether explanations that are yielded from different experiments will result in an integrated pluralism (as advocated prominently by Mitchell 2009) becomes second to questioning the very conditions for identifying disease and translating knowledge based on site-specific identification practices across different domains of medical research and practice.

9.3.2 Scoping Methods

There have been some suggestions to apply a relational account to capture the interdependencies between world, data, data models, and theory in the life sciences (e.g., Leonelli 2019). Scope validity takes this route even further: it side-steps general metaphysical assumptions about diseases though acknowledging that different practices of diagnosing and defining disease come with ontological implications. This perspective urges us to ask in every case study and in every context how exactly the target object is framed and what strategies are applied to evaluate in how far a given research setting conditions the scope of application. The relational approach is well suited to make differences in scientific practices, theoretical assumptions, value judgements, and interests of various actor groups explicit. A relational approach is well-compatible with approaching disease as historical and practical concepts (see e.g., van der Linden and Schermer, Chap. 19, this volume, Binney, Chap. 7, this volume, Fangerau, Chap. 3, this volume). However, it puts a lot of normative weight on the assessment of the *adequate* identification of the target in local settings as well as the evaluation of their matching across contexts of experimentation and application. The relational perspective thus allows us to ask new philosophical questions regarding scientific methods. From a philosophical point of view, we can ponder the dimensions of adequacy in medical research. From a scientific point of view, we can probe methods to assess and increase the scope validity of a model or test regarding its intended use.

From a philosophy of science in practice perspective, it can be a useful first step to turn the attention to methods that are already used by scientists and that can be related to scope validity in some way or another — though they have so far not been analyzed as responding to the same meta-methodological issue. Examples could include backward or reverse translation of animal models and retrospective epidemiological studies. Reverse translation denotes the testing of an intervention that is known to work (or not to work) in humans in animal models. They aim to check for and characterize failures of animal studies to reproduce the human effects, to refine measures of outcome parameters in animals, and/or to compare the validity of different animal models with respect to their capacity of mimicking the proven (in-)effectiveness in human trials. Indeed, good candidates for *scoping methods* seem to be connected to discussions of dissatisfaction with the current structure and practice of biomedical research, and attempts to remedy the experienced shortcomings. For instance, Bert 't Hart and colleagues (2018) define reverse translation as, "when a promising new treatment fails to show efficacy in clinical trials, the reason(s) for failure are investigated by retesting in a relevant animal model (clinic to lab)." They describe reverse translation as an important step to better understand species- (or strain-)specific pathophysiological mechanisms of a disease and problems that result thereof for extrapolation. The scientists echo the complaint from Belzung and Lemoine (2011: 1) that too little research has been funded that applies "the back-translational approach. . . . going from the bedside to the bench." Experimental designs that employ reverse translation could be used as "a learning principle" ('t Hart et al. 2018: 267) for drawing conclusions about the pathophysiological mechanisms that led to a failure of extrapolation from animals to humans. I propose that reverse trials could be used as a scoping method to investigate which target a model might best fit and how to improve and assess the matching.

Conceptualized in this way, we can compare methodological strategies like backward-trials with methods from other subdisciplines such as epidemiology. I mentioned above that some epidemiologists have been very critical of the new biomarker-based research framework to investigate Alzheimer's disease. This framework builds on evidence from longitudinal studies that identified a population according to their performance in neurocognitive tests, and then followed this cohort of people with 'mild cognitive impairment' over years to ascertain their heightened risk to develop symptoms of dementia due to Alzheimer's disease. In contrast, the skeptical epidemiologists argue that evidence from retrospective studies have not been taken seriously enough (Amieva, Glymour, personal communication). In these studies, the starting point is not a putative risk population, but people who have already developed severe symptoms of dementia and received a clinical diagnosis. The epidemiologists then backtrack the patients' medical (and biographical) records for commonalities in their midlife, years before they received the diagnosis. Such studies have shown that a low body mass index was significantly correlated with a dementia diagnosis two decades later (e.g., Qizilbash et al. 2015). This method does not mean to discard neurocognitive testing as tool to identify a population at risk, but it would help to quantify how many patients who receive a dementia diagnosis at the end of the study were overlooked by neurocognitive testing, because cognitive

problems did not occur as early signs in their cases. Again, such retrospective study designs have mostly been discussed as testing a hypothesis (here: the falsification of a potential correlation between obesity and dementia) or to generate alternative aetiological hypotheses (that high metabolic rates might be involved in dementia development) for additional testing in longitudinal epidemiological or laboratory studies. However, we could also analyze and apply such retrospective studies as a *scoping method* for testing what proportion of the clinical diagnoses of Alzheimer's disease did not previously fall into the category of mild cognitive impairment (this has been suggested to me by epidemiologist Hélène Amieva).

9.4 Conclusion

In this paper, I have shown that traditional validity concepts assess the informativeness of a model or test by deploying some abstract concept of the target of modelling. These validity concepts have proven to be ill-suited to assess and refine how well the target of experimentation matches the target of application, not in general, but within the particular local context. To fill this gap, I have introduced a relational approach to medical objects of inquiry that side-steps metaphysical questions about what disease is in general and that is apt for investigating how knowledge generation within one biomedical context conditions the way in which a medical problem is identified in another context. Scope validity does not contradict other validity measures but elucidates a dark spot, and thus could be used complementary to other validity concepts. In contrast to variants of construct validity that assess how close the test hits the abstract entity, scope validity captures how well the target in the experimental test fits to the target in the application setting. The process of forming an ideal of a given disease entity in modern medical sciences (Rosenberg 2002) puts medical scientists and philosophers in the position of having to judge the right way to delineate disease(s). My alternative, relational approach focuses instead on the differences between the relata of animal models, research populations, and the group of people who receive a diagnosis in general health care. Instead of prioritizing one way of delineating disease according to an assessment of how close the given operationalization comes to the idealized disease entity, scope validity addresses the matching (and mismatching) of identifying disease types across concrete contexts.

This tasks researchers in philosophy and science to identify, first, which approaches should be legitimately included in such an analysis; second, how to assess practices of identifying disease within a given context; third, how to examine their matching; and fourth, how to guide the assessment of the matching towards the values we want to see instantiated in a good health care system. Each of these steps, and perhaps the last one most of all, will undoubtedly raise many discussions and concerns, because it might lead us to question the freedom and disinterestedness of scientific inquiry. It is important to raise these (and other) concerns and to examine in detail under what conditions they are warranted. However, it is equally important to

keep in mind that when animal scientists ask for funding from medical research organizations and when pharmaceutical companies run experiments on human beings, there needs to be some accountability for how this research can benefit humans. The resort to abstract concepts of disease has at least in some cases deprivileged attempts to improve scope validity, as exemplified in the case of a pharmaceutical company's strategic choice to not test a potential antidepressant in a better matching, but much smaller subpopulation. If there was a regulatory requirement to assess (and publish this assessment of) the scope validity of a clinical trial design in relation to the animal models that were used to provide mechanistic evidence for a drug's mode of action, such strategic choices would at least be more difficult to advocate. Scientists from various subdisciplines have already developed methods, such as reverse translation, that could be used to examine the matching of scopes across research contexts. Scope validity can serve as a meta-methodological category for identifying, collecting, comparing, and analyzing such *scoping methods*, thereby bringing attention to the epistemological work done in these subdisciplines.

My philosophical account of a relational approach to medical issues and the focus on scientific methods of assessing adequacy bears certain assumptions and limitations with regard to using scope validity as a conceptual tool. It does not provide a fixed set of criteria of adequacy that can serve as a universal standard for evaluating medical research. Rather, the next step would be to provide a more nuanced vocabulary for weighing the premises of local operationalizations of disease within a given experimental design against its intended scope of application. An assessment of a research trial's scope validity neither privileges a certain definition of disease, nor does it necessarily entail that only research should be funded that fits best to received diagnostic criteria. This means that researchers who detail their experiment's or model's scope validity need to question what their disease operationalization implies for the use of their research results in other contexts. Given the social organization of biomedical research as a highly segregated, multi-professional enterprise, the answers to the question of how the premises of research designs in different labs, clinical studies, and application contexts fit to each other will remain underdetermined in many cases. There are too many variables in the process of translating research. I want to argue that this should not be seen as a shortcoming, but rather as an indicator that the concept of scope validity might indeed be of use as a tool for science. The assessment of scope validity will generate questions, which can be made productive when directing philosophical and scientific research into applying and possibly inventing or improving methods to better qualify the adequacy of translational medical research.

Acknowledgement I would like to thank the organizers, my commentator Frank Wolters, and the participants at the workshop on "health and disease as practical concepts" at Erasmus University in Rotterdam, as well as the participants of the Philosophical Club at Bielefeld University, and the commentators and organizer of the Scope Validity in Medicine Panel of the American Philosophical Association's Meeting in Philadelphia 2020 for reading earlier drafts of this paper and giving me very helpful input for my revisions.

References

Abramson, Lyn Y., and Martin E.P. Seligman. 1977. Modeling psychopathology in the laboratory: History and rationale. In *Psychopathology: Experimental models*, ed. J.D. Maser and M.E.P. Seligman, 1–26. San Francisco: WH Freeman.

Alexandrova, Anna, and Daniel M. Haybron. 2016. Is construct validation valid? *Philosophy of Science* 83 (5): 1098–1109.

Ankeny, Rachel A., and Sabina Leonelli. 2011. What's so special about model organisms? *Studies in History and Philosophy of Science Part A* 42 (2): 313–323. https://doi.org/10.1016/j.shpsa.2010.11.039.

Ashe, Karen H., and Kathleen R. Zahs. 2010. Probing the biology of Alzheimer's disease in mice. *Neuron* 66 (5): 631–645. https://doi.org/10.1016/j.neuron.2010.04.031.

Bechtoldt, H.P. 1951. Selection. In *Handbook of experimental psychology*, ed. S. S. Stevens, 1237–1267. New York: Wiley.

Belzung, Catherine. 2013. Innovative drugs to treat depression: Did animal models fail to be predictive or did clinical trials fail to detect effects? *Neuropsychopharmacology* 39 (5): 1041–1051. https://doi.org/10.1038/npp.2013.342.

Belzung, Catherine, and Maël Lemoine. 2011. Criteria of validity for animal models of psychiatric disorders: Focus on anxiety disorders and depression. *Biology of Mood & Anxiety Disorders* 1 (1): 1–16. https://doi.org/10.1186/2045-5380-1-9.

Bertolaso, Marta. 2016. *Philosophy of cancer: A dynamic and relational view*. Dordrecht: Springer.

Burian, Richard M. 1993. How the choice of experimental organism matters: Epistemological reflections on an aspect of biological practice. *Journal of the History of Biology* 26 (2): 351–367. https://doi.org/10.1007/bf01061974.

Carrier, Martin. 2004. Knowledge and control: On the bearing of epistemic values in applied science. In *Science, values, and objectivity*, ed. P. Machamer and G. Wolter, 275–293. Pittsburgh: University of Pittsburgh Press.

Cartwright, Nancy. 2009. Evidence-based policy: What's to be done about relevance? *Philosophical Studies* 143 (1): 127–136. https://doi.org/10.1007/s11098-008-9311-4.

Cook, Thomas D., and Donald Thomas Campbell. 1976. The design and conduct of true experiments and quasiexperiments in field settings. In *Handbook of industrial and organizational psychology*, ed. M.D. Dunnette, 223–326. Chicago: Rand McNally.

———. 1979. *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

Cronbach, Lee J., and Paul E. Meehl. 1955. Construct validity in psychological tests. *Psychological Bulletin* 52 (4): 281–302. https://doi.org/10.1037/h0040957.

Daly, Tim and Lara Keuck. 2024. Alzheimer's Disease. Engaging with an unstable category. In *Handbook of the Philosophy of Medicine*, 2nd Edition. Dordrecht: Springer.

Demazeux, Steeves, and Lara Keuck. 2023. Comment peut-on être précis les yeux fermés? In *Promesses et limites de la psychiatrie de precision*, ed. C. Gauld, E. Giroux, and S. Demazeux, 201–230. Paris: Hermann.

Douglas, Heather E. 2009. *Science, policy, and the value-free ideal*. Pittsburgh: University of Pittsburgh Press.

Feest, Uljana. 2011. What exactly is stabilized when phenomena are stabilized? *Synthese* 182 (1): 57–71. https://doi.org/10.1007/s11229-009-9616-7.

———. 2019. Why replication is overrated. *Philosophy of Science* 86 (5): 895–905.

Germain, Pierre-Luc. 2014. From replica to instruments: Animal models in biomedical research. *History and Philosophy of the Life Sciences* 36 (1): 114–128. https://doi.org/10.1007/s40656-014-0007-0.

Glymour, Medellena Maria, Adam Mark Brickman, Mika Kivimaki, Elizabeth Rose Mayeda, Geneviève Chêne, Carole Dufouil, and Jennifer Jaie Manly. 2018. Will biomarker-based diagnosis of Alzheimer's disease maximize scientific progress? Evaluating proposed diagnostic

criteria. *European Journal of Epidemiology* 33 (7): 607–612. https://doi.org/10.1007/s10654-018-0418-4.

Green, Sara. 2024. *Animal Models of Human Disease*. Cambridge: Cambridge University Press.

Guala, Francesco. 2003. Experimental localism and external validity. *Philosophy of Science* 70 (5): 1195–1205. https://doi.org/10.1086/377400.

Guessoum, Sélim Benjamin, Yann Le Strat, Caroline Dubertret, and Jasmina Mallet. 2020. A transnosographic approach of negative symptoms pathophysiology in schizophrenia and depressive disorders. *Progress in Neuro-Psychopharmacology and Biological Psychiatry* 99: 109862.

Hacking, Ian. 2007. Kinds of people: Moving targets. *Proceedings of the British Academy* 151: 285–318.

Hájek, Alan. 2007. The reference class problem is your problem too. *Synthese* 156 (3): 563–585. https://doi.org/10.1007/s11229-006-9138-5.

't Hart, Bert A., Jon D. Laman, and Yolanda S. Kap. 2018. Reverse translation for assessment of confidence in animal models of multiple sclerosis for drug discovery. *Clinical Pharmacology & Therapeutics* 103 (2): 262–270. https://doi.org/10.1002/cpt.801.

Hauswald, Rico, and Lara Keuck. 2017. Indeterminacy in medical classification: On continuity, uncertainty, and vagueness. In *vagueness in psychiatry*, ed. Geert Keil, Lara Keuck, and Rico Hauswald, 93–116. Oxford: Oxford University Press.

Herzog, David P., Holger Beckmann, Klaus Lieb, Soojin Ryu, and Marianne B. Müller. 2018. Understanding and predicting antidepressant response: Using animal models to move toward precision psychiatry. *Frontiers in Psychiatry* 9. https://doi.org/10.3389/fpsyt.2018.00512.

Hsiao, Karen, Paul Chapman, Steven Nilsen, Chris Eckman, Yasuo Harigaya, Steven Younkin, Fusheng Yang, and Greg Cole. 1996. Correlative memory deficits, a elevation, and Amyloid Plaques in transgenic mice. *Science* 274 (5284): 99–103. https://doi.org/10.1126/science.274.5284.99.

Huber, Lara, and Lara Keuck. 2013. Mutant mice: Experimental organisms as materialised models in biomedicine. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 44 (3): 385–391. https://doi.org/10.1016/j.shpsc.2013.03.001.

———. 2017. Philosophie der biomedizinischen Wissenschaften. In *Grundriss Wissenschaftsphilosophie: Die Philosophien der Einzelwissenschaften*, ed. Simon Lohse and Thomas Reydon, 287–318. Hamburg: Meiner Verlag.

Jack, Clifford R., David A. Bennett, Kaj Blennow, Maria C. Carrillo, Billy Dunn, Samantha Budd Haeberlein, David M. Holtzman, et al. 2018. NIA-AA research framework: Toward a biological definition of Alzheimer's disease. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association* 14 (4): 535–562. https://doi.org/10.1016/j.jalz.2018.02.018.

Keller, Evelyn Fox. 2000. Models of and models for: Theory and practice in contemporary biology. *Philosophy of Science* 67 (September): 72–86. https://doi.org/10.1086/392810.

Kelley, Truman Lee. 1927. *Interpretation of educational measurements*. Yonkers/Chicago: World Book Company.

Kendler, Kenneth S., and Josef Parnas. 2012. *Philosophical issues in psychiatry. II, Nosology*. Oxford: Oxford University Press.

Keuck, Lara. 2020. A window to act: Revisiting the conceptual foundations of Alzheimer's disease in dementia prevention. In *Preventing old age and decline? Critical observations on aging and dementia* ed. A. Leibing, S. Schicktanz. New York, Oxford: Berghahn Publishers, 19–39.

Kitcher, Philip. 2003. *Science, truth, and democracy*. New York: Oxford University Press.

Kutschenko (=Keuck), Lara. 2011a. In quest of 'Good' medical classification systems. *Medicine Studies* 3 (1): 53–70. https://doi.org/10.1007/s12376-011-0065-5.

———. 2011b. How to make sense of broadly applied medical classification systems: Introducing epistemic Hubs. *History and Philosophy of the Life Sciences* 33 (4): 583–602.

Lemoine, Maël. 2011. *La désunité de la médecine: Essai sur les valeurs explicatives de la science médicale*. Paris: Hermann, Impr.

Leonelli, Sabina. 2019. What distinguishes data from models? *European Journal for Philosophy of Science* 9: 22. https://doi.org/10.1007/s13194-018-0246-0.

Longino, Helen. 2002. *The fate of knowledge*. Princeton: Princeton University Press.

Markus, Keith A., and Denny Borsboom. 2013. *Frontiers of test validity theory: Measurement, causation and meaning*. New York: Routledge.

McKinney, William T., and William E. Bunney. 1969. Animal models of depression. *Archives of General Psychiatry* 127: 240–248.

Meunier, Robert. 2012. Stages in the development of a model organism as a platform for mechanistic models in developmental biology: Zebrafish, 1970–2000. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 43 (2): 522–531. https://doi.org/10.1016/j.shpsc.2011.11.013.

Mitchell, Sandra D. 2009. *Unsimple truths: Science, complexity, and policy*. Chicago: University of Chicago Press.

Mol, Annemarie. 2002. *The body multiple: Ontology in medical practice*. Durham: Duke University Press.

Mullan, Mike, Fiona Crawford, Karin Axelman, Henry Houlden, Lena Lilius, Bengt Winblad, and Lars Lannfelt. 1992. A pathogenic mutation for probable Alzheimer's disease in the APP gene at the N–Terminus of β–Amyloid. *Nature Genetics* 1 (5): 345–347.

Parker, Wendy S. 2020. Model evaluation: An adequacy-for-purpose view. *Philosophy of Science* 87 (3): 457–477.

Plutynski, Anya. 2018. Review of philosophy of cancer: A dynamic and relational view, by Marta Bertolaso. *History and Philosophy of the Life Sciences* 40: 1. https://doi.org/10.1007/s40656-017-0165-y.

Qizilbash, Nawab, John Gregson, Michelle E. Johnson, Neil Pearce, Ian Douglas, Kevin Wing, Stephen J.W. Evans, and Stuart J. Pocock. 2015. BMI and risk of dementia in two million people over two decades: A retrospective cohort study. *The Lancet Diabetes & Endocrinology* 3 (6): 431–436. https://doi.org/10.1016/s2213-8587(15)00033-9.

Rheinberger, Hans-Jörg. 2006a. Vers la fin des organismes modèles? In *Les organismes modèles dans la recherche médicale*, ed. G. Gachelin, 275–277. Paris: Presses Universitaires de France.

———. 2006b. Gaston Bachelard und der Begriff der ‚Phänomenotechnik'. In *Epistemologie des Konkreten: Studien zur Geschichte der modernen Biologie*, 37–54. Frankfurt am Main: Suhrkamp.

Robins, Eli, and Samuel B. Guze. 1970. Establishment of diagnostic validity in psychiatric illness: Its application to Schizophrenia. *American Journal of Psychiatry* 126 (7): 983–987. https://doi.org/10.1176/ajp.126.7.983.

Rosenberg, Charles E. 2002. The Tyranny of diagnosis: Specific entities and individual experience. *The Milbank Quarterly* 80 (2): 237–260. https://doi.org/10.1111/1468-0009.t01-1-00003.

Schaffner, Kenneth F. 1998. Genes, behavior, and developmental emergentism: One process, indivisible? *Philosophy of Science* 65 (2): 209–252. https://doi.org/10.1086/392635.

———. 2012. A philosophical overview of the problems of validity for psychiatric disorders. In *Philosophical issues in psychiatry II: Nosology*, ed. Kenneth S. Kendler and Josef Parnas, 169–189. Oxford: Oxford University Press.

Schaffner, Kenneth F. Forthcoming. Construct validity in psychology and psychiatry. (Draft version Oct 2018; available from the author on request at kfs@pitt.edu).

Solomon, Miriam. 2015. *Making medical knowledge*. Oxford: Oxford University Press.

———. 2022. On validators for psychiatric categories. *Philosophy of Medicine* 3 (1): 1–23.

Steel, Daniel. 2008. *Across the boundaries: Extrapolation in biology and social science*. Oxford: Oxford University Press.

Strzebonska, Karolina, and Marcin Waligora. 2019. Umbrella and basket trials in oncology: Ethical challenges. *BMC Medical Ethics* 20: 58. https://doi.org/10.1186/s12910-019-0395-5.

Sullivan, Jacqueline A. 2009. The multiplicity of experimental protocols: A challenge to reductionist and non-reductionist models of the unity of neuroscience. *Synthese* 167 (3): 511–539. https://doi.org/10.1007/s11229-008-9389-4.

Tabb, Kathryn. 2017. Philosophy of psychiatry after diagnostic kinds. *Synthese* 196 (6): 2177–2195. https://doi.org/10.1007/s11229-017-1659-6.

Tabb, Kathryn, and Maël Lemoine. 2021. The prospects of precision psychiatry. *Theoretical Medicine and Bioethics* 42: 193–210. https://doi.org/10.1007/s11017-022-09558-3.

Wagenknecht, Susann, Nancy J. Nersessian, and Hanne Andersen. 2015. *Empirical philosophy of science: Introducing qualitative methods into philosophy of science*. Dorndrecht: Springer.

Willner, Paul. 1984. The validity of animal models of depression. *Psychopharmacology* 83 (1): 1–16. https://doi.org/10.1007/bf00427414.

———. 1994. Animal models of depression. In *Handbook of depression and anxiety. A biological approach*, ed. Siegfried Kasper, Johan A. den Boer, and J.M. Ad Sitsen, 291–316. New York: Marcel Dekker.

Willner, Paul, and Paul John Mitchell. 2002. Animal models of depression: A diathesis-stress approach. In *Textbook of Biological Psychiatry*, ed. H. D'Haenen, Johan A. den Boer, and Paul Willner, 703–706. Chichester: Wiley.