Chapter 29

Psychiatric classification: An a-reductionist perspective

Jan-Willem Romeijn, PhD

Hanna van Loo, MD, PhD

University of Groningen

1. Introduction

Many psychiatric disorders capture heterogeneous classes of patients, in terms of their aetiology, course of illness, and response to treatment (Baumeister 2012). For instance, of all patients with a first episode of major depression, about one third will have only one lifetime episode, whereas two thirds will suffer from recurrent or chronic episodes (Eaton 2008). This variation in patients with the same diagnosis hampers prediction and treatment assignments in clinical practice. It has evoked continuing efforts to improve psychiatric disease classification such as the Diagnostic and Statistical Manual of Mental Disorders and International Classification of Disease (cf. Kupfer 2008).

Recently, the American Psychiatric Association has invited researchers world-wide to contribute to improving the DSM, and propose alternative disease or subtype definitions based on empirical evidence according to the so-called "empirically-driven continuous improvement model" (First 2017, Kendler 2013). Several validators – e.g., familial aggregation, biological markers, course of illness, response to treatment – have been proposed as criteria to judge whether new proposals will improve on current disease definitions (Kendler 2013). If there is clear evidence that alternative definitions outperform the current ones in terms of validity, reliability, or clinical utility, this might lead to a change in specific diagnostic categories (First 2017).

Note that performance is put in terms of validity, reliability and clinical utility. All of these, we note, relate to the more general goals of accurate prediction and effective intervention. Consider the afore-mentioned validators. Bio-markers and familial aggregation signal robust predictive properties, and both expected course of illness and response to treatment relate to therapeutic interventions. And much like the measure of validity, the measures of reliability and

clinical utility relate to prediction and intervention: the reason to aim for them is that it is desirable that different clinicians have similar expectations, similar views on what is wrong with the patient, and similar ideas on what can best be done. Therefore, without suggesting in any way that these goals exhaust the purposes of psychiatric classification, we maintain that classifications are generally better when they lead to improved opportunities for prediction and intervention (cf. Cartwright and Hardie 2012). This will be true for clinicians, researchers, and patients, although all these stakeholders will have differing predictive concerns.[1]

But how to identify classes of patients that improve prediction and intervention, and outperform the existing classifications? The focus of this chapter will be how statistical methods can be utilized to contribute to this goal.[2] Our outlook on the improvement of psychiatric classification, moreover, will offer insight into the fact that classification schemes can include characteristics that pertain to entirely different theoretical domains, or explanatory levels. Is it problematic when a classification scheme sorts individuals by means of specific pathophysiological characteristics, as well as by psychological and social ones? Towards the end of this chapter we will answer this question in the negative, and motivate our position.

[1] One possible criticism of viewing psychiatric classification this way is that it ignores one of its most important purposes. A classification may also support an understanding of disorders, an explanation of a course of events, and thereby an instrument for therapists and a consolation for those who suffer from mental disorder. For reseachers understanding is similarly important, even if they will have a different use for it. For now, we want to suggest that the value of understanding and explanation is often in the empirical consequences that they have. Insight into a disorder helps us to apply our ideas more adequately and reliably, to gain a sense of control, to support our decisions in dealing with the disorder. Therefore, to some degree we view the explanatory value of a classification as derivative.

[2] While there are certainly other possible approaches to psychiatry, e.g., medical casuistics or single-case research, current medical research is dominated by the practice of statistics, i.e., the collection of data and the use of statistical methods to investigate mental illness. In this chapter we approach psychiatric research from this statistical angle.

This chapter is structured as follows. First we will show that the problem of finding the right categories is equivalent to the so-called "reference class problem", an essentially statistical problem well-known from the philosophy of science literature (section 2). Then we discuss how model construction and selection and causal modeling methods can be used to identify adequate classes of individuals (sections 3 and 4). Next we argue that these methods promote a so-called a-reductionist perspective towards the variety of explanatory levels in psychiatry (section 5). We conclude by giving a brief summary of what we have claimed (section 6).

2. Psychiatric classification as a reference class problem

We introduce a statistical perspective on the problem of psychiatric classification and then show how it coincides with the problem of the reference class[3]. Towards the end we look ahead and indicate how our perspective links up with a particular view on explanatory levels.

Classification as statistical model building

A classification scheme generates a grouping of the population based on certain characteristics assigned to the individuals. Depending on the granularity of the classification, the groups will be larger or smaller. For a system that only contains two binary variables, for instance "depression yes/no" and "psychosis yes/no", there will be 2×2, hence 4 groups. For a very rich classification scheme, on the other hand, every individual might be contained in their own group. Of course, the usual disease classifications will have a granularity that sits in between such extremes, and that is endowed with further structure. The DSM-5, for instance, specifies a grouping into a large number of disorders, into the hundreds, by means of a collection

[3] To avoid terminological confusion: the notion of reference class here is different from the reference class in analyses of categorical variables.

of even more symptoms. These characterisations in terms of symptoms allow us to define very many subgroups within these disorders, of which some will be clinically meaningful while others are not (cf. Olbert et al 2014).

In general, the problem of classification is that of finding the relevant set of characteristics of individuals. Viewed in this way, classification sits close to the core tasks of psychiatric scientists. Finding out about the factors that matter to the prediction of, and intervention on psychiatric disorders encompasses much of their research. It involves measuring, constructing and then selecting the right variables, and determining relations among them by means of experimental and observational studies. In the philosophy of science this is often referred to as the construction and evaluation of "models" (cf. Morgan and Morrison 1999). Several of the other chapters in this volume are directly or indirectly concerned with models. The construction of a model is at stake in, for instance, "designing control panels" by which we judge a clinical course of action (Campbell [this volume]), pitching the patient descriptions at the right level (Pine [this volume]), and even in crafting new concepts in the psycho-social realm, by means of which we can capture the experiences or raw empirical facts of a clinical practice (Parnas, Gallagher [this volume]).

A key problem in psychiatric classification is that of identifying homogeneous and maximally distinct groups. We look for interclass heterogeneity, i.e., classes that are dissimilar on variables of interest, and intraclass homogeneity, i.e., classes of similar individuals with respect to these variables. As said, in this chapter we consider the role of statistical methods in the task of classification, and we take classification to have the purpose of supporting predictions and interventions. Accordingly, what it means for patients in the classes to be similar, and between classes to be dissimilar, is that they are alike and different in terms of the characteristics

salient for what we want to predict and control. There is intra-class homogeneity, and inter-class heterogeneity, when within the classes we find little variation among characteristics that matter for those goals, and when between the classes we find large variation.

It is extremely rare that membership of some class, i.e., having a certain combination of characteristics, fully determines a particular course of illness, or a particular result of an intervention. The normal situation is that a class is associated with chances, not certainties. In this statistical context, intra-class homogeneity, and hence the similarity of members of the class, means that all the members have roughly the same chance for some event or result. Inter-class heterogeneity, in the same vein, means that those chances vary widely when moving from one class to the next. For a homogeneous class, then, the proportions within the class will be good estimations of the chances for the members of the class. A classification that satisfies intra-class homogeneity thus identifies groups for which we can build up statistical knowledge, by observing proportions within the group and taking them as estimates for chances that then apply to the individuals in that group. This is arguably the core idea of statistics.

The problem of the reference class

In the philosophy of science, groups on which we can base chance ascriptions are termed "reference classes" (Reichenbach 1949, Hajek 2007). A useful reference class for an individual is a group to which that individual belongs such that we can infer stable chances for the individual from the observed proportions of the group. Say that we offer a certain treatment to all individuals labelled with a disorder from our classification scheme, record the proportion of recoveries within that group, and then take the proportion as indicative of the chance of recovery for any person suffering from the disorder. If the classification scheme groups the individuals

together in the right way, the chance ascriptions to individuals will be predictively accurate, and useful in making the intervention decisions. The problem that we are facing when classifying mental disorders is precisely the problem of the reference class: what individuals shall we group together for the purpose of determining these chances?

By analogy, say that we are asked to sort a large collection of dice, with a varying number of sides, of which an unknown number of sides shows a 1. Note that this need not only be six-sided dice, and that the numbering on the dice might have duplications, e.g., when several sides all show a 1; see figure 1.

Figure 1. A collection of dice with different numbers of sides.

Now imagine that our aim is to predict whether we will roll a 1 with a randomly selected die. We then do best to disregard the colour and weight of the dice, and focus only on the number of sides showing 1 and on the total number of sides when we make a grouping. Further, if we are asked to make groups, we might decide to isolate a set of dice for which rolling a 1 has a low chance, one for which the chance is high. Depending on the collection of dice given to us, we might find that there really are two distinct sets of dice, one with dice that have between ⅔ and ¾ of the sides marked 1, and another that do not have any 1's, or else a single 1 among at least ten sides. Such sets show high intra-class homogeneity, with little variation in the chances for each individual die within the set, and high inter-class heterogeneity, because the difference

between the chances of rolling a 1 between the two sets is large. But we might be less lucky with our dice collection, with the chances on a 1 spread over the whole spectrum between 0 and 1 without any sort of grouping. However this may be, we will group the dice according to the predictive goals we set ourselves, and we will focus on characteristics that are relevant for determining the salient chances.

Next to this example with dice, consider a concrete psychiatric example that is structurally the same. Say we want to determine the intensity of monitoring for an individual patient who just recovered from a first episode of major depression. Then it would be useful to know the probability of this patient on long-term remission versus chronic or recurrent episodes of depression. Suppose that, in general, for patients who recovered from a first episode of depression, this chance is roughly 50%. Could we imagine a better reference class for this individual patient? An good reference class would be a sub-class of patients recovered from a first episode of MD that are similar, in such a way that we can infer the probability for this individual patient from the group average; we want this chance ascription to be stable. Ideally, chances in this reference class would be extreme, i.e., very low or very high. The percentage for all patients recovered from a first episode of MD, to wit, 50%, is not very informative.[4]

From the example it will be apparent what it means for a grouping, and hence for a model or a classification scheme, to generate useful reference classes. The classification of the individuals must facilitate accurate predictions of, and effective interventions on the phenomena that we care about. Optimizing a psychiatric classification scheme is, at least in part, that kind of

[4] Readers with philosophical inclinations may wonder if chances for the individual can make sense at all, but ideas from the philosophy of science can help us ground the requisite notion of chance conceptually. By employing ideas from emergentism and multiple realizability, we can overcome reductionist challenges to the coherence of single-case chances, including chances assigned to variables and events that are characterized at a high-level of description.

exercise: it concerns the selection of criteria for the formation of groups that can serve as reference classes for stable and distinct chance ascriptions to variables of interest, either for the purpose of accurate predictions or for effective intervention. Importantly, this is an empirical issue: we determine the groupings not on the basis of some preconceived notion of natural kind, or on the basis of a preconceived explanatory level, but primarily by the empirical facts of which characteristics facilitate prediction and intervention.[5]

Looking ahead

The remainder of this chapter is devoted to working out some of the consequences of viewing nosological reform in this way, namely, as ultimately a statistical affair of forming groupings based on the characteristics of the individuals (cf. Grove and Meehl 1996). The next two sections point to specific statistical methods that may help us to identify groupings with relevantly similar individuals, for which distinct and stable chances can be determined. In the section following that, we consider the more theoretical implications of our perspective on nosological reform, in particular the non-committal position in the reductionism debate that is entailed by it. We can already note that nothing in the foregoing suggests that we have to limit our search for salient characteristics to a specific explanatory level, for instance by looking only to neurological variables, or cognitive and social ones. Any characteristic of an individual is in principle suitable for inclusion into the classification scheme, and they can all be treated on a par.

[5] This empirically-driven way of classifying individuals is reminiscent of Hathaway and McKinley (1940) and the development of the Minnesota Multiphasic Personality Inventory (MMPI), a standardized psychometric test of adult personality and psychopathology. It also reminds of Meehl (1956), who discusses the MMPI extensively. We thank Peter Zachar and Marcus Eronen for pointing us to these respective parallels.

In what follows, we will keep returning to this a-reductionist implication of our view on classification.

3. Models: construction and selection

In this section we devote our attention to the two research phases of model construction and model selection, because they hold particular promise for the design of classifications. Moreover, as we argue at the end, the statistical tools that help us construct and select models are neutral towards explanatory levels, and therefore support the afore-mentioned position of a-reductionism.

Statistical methods for classification design

What statistical methods can be used to contribute to the design of such classification schemes? For one, ordinary statistical analysis, carried out against the backdrop of a model, can be highly instrumental: hypothesis tests, parameter estimations, and statistical inferences may all contribute to the design of classifications, e.g., by determining the relative importance of characteristics that are taken into consideration. However, we also find methods that are suited to the task in the research phase that precedes statistical analysis, namely in the construction of a statistical model, and in the phase that follows it, namely in the evaluation of those models. We concentrate now on these methods.

On the side of model construction, certain statistical learning methods can be used to discern similarity patterns in characteristics of patients that were not known beforehand (Lubke and Muthen 2005, Hastie 2013), and thereby suggest specific classifications. The advantage of

using statistical learning methods is that these can evaluate vast numbers of patient characteristics in large samples of subjects. Based on similarity patterns in these patient characteristics, these methods can divide subjects into subgroups with high intraclass homogeneity, and pronounced inter-class differences. The resulting statistical models might underpin alternative classifications.

A good example comes from the research into more homogenous subtypes for depression (cf. Baumeister 2012). How can we use statistical methods to improve on current subtypes, such as the traditional division into melancholic and atypical depression (American Psychiatric Association 2013)? In a recent study, we used data of the World Mental Health Survey (Kessler 2008) of more than 8,000 subjects with a lifetime depressive episode (van Loo 2014, Wardenaar 2014). These subjects were interviewed about a range of clinical characteristics, such as their symptoms during the depressive episode, their age when they became first depressed, psychiatric comorbid disorders, and whether their parents also suffered from depression. The subjects also reported on the course of their depressive illness, i.e. on the number of depressive episodes they had, the chronicity of these episodes, whether they were ever hospitalized for depression, and whether they were disabled to work.

Statistical learning methods were then deployed to discover classes of patients with similar course of illness patterns, based on these clinical characteristics. Using penalized regression methods, we constructed a variety of models, ranging from more to less complex in terms of the numbers of included clinical characteristics. After a phase of model construction and statistical analysis, we assessed the performance of the models by means of cross-validation. We selected the model that best predicted the course of illness, and defined three subtypes of depression, with a low, intermediate and high risk for a severe course of illness. When we tested

the accuracy of this model in new data, the proposed subtypes indeed differentiated between subjects with a more severe, intermediate or mild course of illness (Figure 2, Kessler 2016).

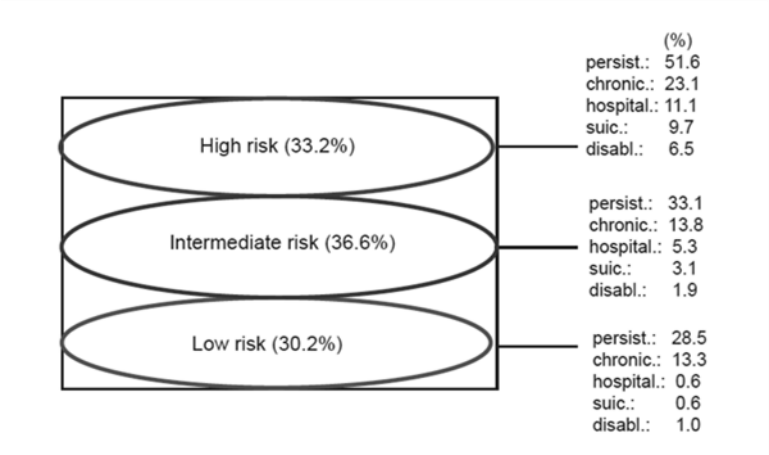

Figure 2: Association of identified risk clusters with course of illness after 10-12 years in 1056 subjects with lifetime depression in the US National Comorbidity Survey (Kessler 2016). This figure presents prospective associations between initial cluster scores (1990-1992) and subsequent persistence and severity of course of depression (2001-2003). The outcomes measured at follow-up concern percentages of years with depressive episodes (persist.; persistence), the episodes lasting most of the year (chronic.; chronicity), hospitalization (hospital.; hospitalization) and suicide attempts since baseline (suic.; suicide attempts), and current disability (disab.; disability), in the NCS data. Area under the curves (AUCs) for the three cluster classification varied between 0.60-0.69 for the outcomes indicating years with (chronic) episodes, and 0.70-0.73 for outcomes indicating severity (hospitalization, suicide attempts, and disability). Kessler 2016

Assessing models: finding the sweet spot

Notice that these previous studies (van Loo 2014, Wardenaar 2014) used cross-validation to select the statistical model that best predicted the course of depression. But cross-validation is only one of the many methods for doing this. Under the header of statistical model selection (Claeskens and Hjort 2008) there is a large literature on how to select variables for inclusion in a

model.[6] For instance, there are so-called information criteria, ICs for short, that provide every model with a score, expressing how well the model fits the data. The Akaike and Bayesian ICs are among the most commonly used ones. Many of these model selection methods target the same measure of model adequacy, namely the expected predictive performance of the model. For the purpose of this chapter, we will bypass the conceptual questions that this might raise, and focus on what this measure of adequacy invariably leads to: a trade-off between the complexity of the model, and the likelihood for the data of the best fitting hypothesis in the model.

A quick illustration will make apparent why this is of particular importance for our purposes. Consider a sample of people suffering from a mental disorder, and two classification schemes or models that may be used to fit the data of the patients. The first of these has very few classes, lumping together large patient groups, while the second has so many of them that every individual occupies their own group. The problem with the first classification is that it will not be sufficiently differentiating, for example between patients with a severe course of illness versus a mild course of illness. It will lead to highly heterogeneous groups, and therefore to inaccurate predictions. The second classification will also be problematic though, albeit for a different reason. If the classification is so fine-grained that everyone will be the sole datum for their own grouping, there is hardly any basis for extrapolating from the observed individuals to other individuals in the population: a single data point might give some information about the class to which that individual belongs, but it is far too little for reliable predictions. The upshot is that the

[6] The suggestion of model "selection" is somewhat misleading because the model selection methods merely evaluate and compare models. The researcher's act of selecting one model is more akin to making a decision than forming a judgment, and therefore something best understood by means of decision theory. It involves more than assessing evidential relations between data and model; it also concerns the utility of the outcome.

predictions stemming from the classification will be unreliable. For either model, the predictive performance will thus be found wanting.

The point of all the model selection methods is to find the sweet spot between these two extremes. On the one hand we want to avoid overfitting, i.e., picking up on noisy or unimportant individual differences in the data and viewing them as signals, thereby identifying too many subgroups. This corresponds to optimizing on inter-class heterogeneity: we want to avoid distinguishing groups that are not all that different. On the other hand we want to avoid underfitting, i.e., failing to pick up on signals in the data because we lack the means to detect genuine differences among the individuals. This corresponds to optimizing on intra-class homogeneity: we do not want to miss out on salient distinctions among patients. Model selection methods help us to make this trade-off in a broadly data-driven way, and thus optimize the expected predictive performance.

Discussion: subject-specific knowledge and explanatory levels

A few comments on these statistical methods are in order. First, there are numerous model discovery and selection methods and they all strike the balance between fit and complexity in a slightly different way. We certainly do not want to suggest that the issue of how to strike the balance can be delegated to a statistics department. Application of these methods is only helpful if it is combined with detailed knowledge about clinical psychiatry, and with a good understanding of the assumptions that underpin the methods. Second, the results of model selection methods will depend strongly on the predictive targets that we set. There is no guarantee that the set of variables that appears to be the sweet spot for predicting the course of one particular illness, will also be the sweet spot for predicting treatment response. It may well

be that we have to maintain several task-specific models. Although we cannot develop this idea in any detail, it is in principle possible to select a set of variables for optimal performance on a range of predictive tasks, simply by finding a compromise between the demands placed by the different predictive goals. How accurate these predictions are, will depend on the nature of the compromise but also on how regular and noisy the phenomena themselves are.

Finally, notice that in the foregoing we did not discuss the level of description of variables that are considered for inclusion in the classification scheme, or the scientific discipline from which they originate. That is simply irrelevant to the application of the model selection methods: all variables, from biological and behavioral to cognitive and social, are treated on a par by the statistical methods under consideration. The methods thus offer a particular grip on the classification of disorders for which multiple explanatory levels are implicated. In practice it may not always be easy to combine data from different levels, e.g., genetic data with data from the cognitive and social realm, for example because there are few large data sets in which all these characteristics are combined. The point we want to make is that the statistical approach to psychiatric classification that is under discussion here does not constrain us to a single level. All that matters is the role of a variable in improving the predictive performance of the classification scheme.

4. Causal modeling

This section considers the role of causal network models, a statistical tool for determining causal relations, in the design of psychiatric classifications. After introducing them in a worked example, we argue for their usefulness and touch on their fit with our a-reductionist viewpoint.

The importance of interventions

Besides predicting events, an important goal of science is to intervene. This is true in particular for medical science with its focus on treatment. A core desideratum for psychiatric classification is that it allows us to intervene on, and change the course of mental disorders for the better: the classification needs to guide treatment decisions. The desideratum for a classification is therefore that individuals from different classes of patients respond to treatment options in the same way, and that for each group of patients there is at least one treatment option that offers a high probability of success. For each individual patient, an optimal treatment can then be determined by referring back to the classification. Accordingly, the classification can then support the organization of treatment programs.

To facilitate maximally effective clinical interventions, it is helpful when our classification meshes with the causal structure of the disorder. Insight into the causal structure will reveal what factors initiate, promote, moderate, mediate, or otherwise modify the disorder, and how we can influence these factors to positive effect. Unfortunately silver bullets are a rarity in a therapeutic context but we might hope that the causal structure among the factors gives us a statistical grip, in the sense that we gain some control over the chances of recovery. For steering nosological reform, a crucial question is therefore whether the classification scheme facilitates interventions with good statistical properties: we want to include variables or characteristics if

they are useful in the specification of treatments that are effective, in the sense that they increase the recovery chances. Importantly, it is not thereby required that we fully expose the mechanisms that are driving the disorders, because we can also gain causal knowledge, and hence control, through derivative variables.

The statistical toolbox of the psychiatric researcher already includes designs and methods that help evaluate treatment: randomized controlled trials, hypothesis testing and parameter estimation, e.g., regression analysis, and various ways of controlling for confounders. However, if the aim is to lay our hands on causal structure, statistics has its well-known shortcomings, as laid down in the slogan "correlation is not causation". It is received wisdom that statistical relations between variables cannot help us establish the causal ties between the underlying events.

Fortunately, the past three decades have seen the development of new statistical methods, developed in statistical science but also in computer science and philosophy, aimed at determining causal structure (Glymour, Spirtes and Scheines 2001). The methods go by the name of causal networks, or sometimes causal Bayesian networks even though many of the statistical methods involved in using these methods are not Bayesian but frequentist. In our view causal networks are underused in the sciences, considering their potential value. As we will argue below, psychiatric classification seems especially suitable for their application.

Causal network models

The key idea of causal networks is that correlations and dependency relations between variables can be captured in a network. The variables are the nodes in such networks and the arrows in between represent statistical dependencies. These arrows are then interpreted causally,

so that the graph helps us to determine what we can expect after we have changed the value of one of the variables. This is what makes causal networks different from other statistical methods that help us to determine the effects of treatments: they offer an explicit grip on the interventions. For our purposes, it suffices to discuss a few of the basic ideas by means of an example. Our goal with this is to illustrate how causal networks can be instrumental to making decisions over the salience of variables, and hence over their inclusion into a classification scheme.[7] For further detail on causal network models, we refer to Glymour et al (2001) and, more accessible, Pearl (2018).

Say that we have done an observational study recording whether or not individuals with panic disorder received treatment with a selective serotonine reuptake inhibitor, here denoted simply as *SSRI*, and also whether or not they reported a recovery after 8 weeks, denoted *RepRec*. Imagine that we found a *negative* correlation between the two, $P(RepRec|SSRI) < P(RepRec)$, i.e., somewhat surprisingly the treatment seems to have a negative effect on recovery. We can now construct a simple network that expresses this correlation, interlinking the variables and tentatively marking the link as negative. Moreover, considering that the treatment event preceded the recovery and interpreting the link as casual, we can orient the relation, as depicted on the left in figure 3.

Figure 3: A simple causal network for treatment with SSRI and self-reported recovery. The observation study suggests a negative connection but the RCT shows a positive impact.

[7] The narrative of this section also illustrates Simpson's paradox (Pearl 2000, chapter 6), which can be illuminated very well by means of causal networks. But the emphasis will not be on the paradox itself.

Imagine that we have also carried out a randomized controlled trial (RCT) on the efficacy of the SSRI compared to placebo treatment in panic disorder. And that, despite the negative correlation in the field study, actively administering the treatment, denoted as *Do*[*SSRI*], to a randomly selected set of individuals has a positive impact on their recovery, as depicted on the right of figure 3. What might explain this seeming inconsistency? The answer to this question is that in the observational study the explaining variable *SSRI* is "confounded": it is correlated with other variables, not present in our narrative thus far, that have an impact on the outcome variable *RepRec*. Doing an RCT allows us to determine the separate impact of SSRI on recovery, by removing the correlations with all other relevant variables, or at least attempting to remove them.

A development of the narrative brings the confounder into view. Say that, in the field study, further characteristics of the population before the treatment were recorded, namely their age, gender, and whether they had comorbid depression, denoted by *DiaDep*. Including age and gender as nodes in the causal network shows no moderation of the negative association between SSRI and *RepRec*, but the variable *DiaDep* does (Figure 4). In the field study, subjects with panic disorder and depression were more often treated with an SSRI, and *DiaDep* was therefore highly correlated with SSRI use. Furthermore, and crucially, the prospects of recovery from panic disorder are much worse if there is comorbid depression (Roy-Byrne 2000). In the observation study the recovery rate for individuals who received SSRI will therefore be lower. This is not because the SSRI treatment itself is detrimental to recovery. It is because the individuals who were given SSRI were by and large those individuals who had comorbid depression symptoms, and who will therefore recover less easily.

INSERT FIGIRE 4 HERE

Figure 4 provides a more complete network among the salient variables on the left, as well as a network corresponding to the RCT on the right. Note the marked difference between merely recording whether or not the SSRI treatment was given, as expressed in the variable *SSRI* on the left, and actively administering the SSRI treatment, which we will denote by the variable Do[*SSRI*] on the right. In the observation study, the positive impact of SSRI on recovery is masked by the negative correlation that is established through the comorbid depression. Administering SSRI irrespective of comorbid depression removes this correlation, so that the positive impact of SSRI on recovery can come to the fore.

The merits of causal networks

The foregoing merely illustrates the idea of causal networks. It shows the benefits of classifying patients with panic disorder into two subgroups, those with and those without comorbid depression. If we do not distinguish the subgroups and work with the coarse-grained classification, we misread field data on the efficacy of SSRI treatment, and this may lead us astray when we are considering whether or not to administer an SSRI. Getting to a more complete causal structure, and adapting the classification accordingly, helps us to determine effective clinical interventions.

The more standard statistical treatment, based on RCT's and the methods for identifying confounders, would seem perfectly fine for using comorbid depression as a means to create subgroups of patients with panic disorder. So what is the use of the causal network models? We think the foregoing illustrates that causal networks are helpful for guiding our thinking about the inclusion of variables into a classification scheme. The example is of course simplistic and

idealized. Psychiatric science offers numerous cases that require much more detail, with larger numbers of variables and substantial uncertainty over the causal connections between them. We can easily expand the networks above, up to the point where our intuitions fail and standard statistical tools become clunky and unclear. The theory of causal networks is a well-developed and complete toolkit for investigating the causal structure of systems of variables, and for determining what difference certain interventions make to the chance of recovery in distinct subgroups of patients. Decisions about the inclusion or not of a variable in a classification can thus be supported by causal modeling.

Another important advantage of causal networks is that, also in much more involved narratives and models, they offer a formally precise grip on interventions. The administering of SSRI, for example, can be represented precisely in terms of an operation on the original network structure, as illustrated in figure 4. This is helpful in at least two ways: we can systematically determine what our current model and its model estimates entail about the results of interventions; and we have a systematic means to adapt our model if our predictions about the results of interventions are not borne out. Specifically, in the example, the first network that linked SSRI to recovery negatively turned out to be at variance with the results of the intervention study. And this led us to search for, and eventually identification of confounding variables. The networks, in short, are convenient tools in the construction of models, and in the derivation of predictions following interventions.

As a further motivation for using this methodology, the theory of causal networks is undergoing rapid development. There is active research on causal methods in statistics, machine learning, and philosophy, and there are many interesting areas of research that deserve mention here. We want to end this section by mentioning a development that we find particularly

important for psychiatric classification. One pressing problem in psychiatry is that it is causally complex. Disorders may be triggered by a multitude of factors, manifesting on different explanatory levels. We are typically presented with a cacophony of inter-dependent variables that are all salient, and amongst which there is no clear order of relative prominence. It becomes virtually impossible to say in general what caused a disorder: everything did to some extent. Recent work on causal feature learning (Chalupka 2016) may provide the start of a solution to this. Machine learning techniques can construct macroscopic variables from large collections of factors, in such a way that these macroscopic variables play the requisite causal role. The method enables us to identify global characteristics of systems, in this case of individual patients, that are causally relevant for the course of illness. We are not yet in the position to apply this machinery to the case at hand, but the idea of causal feature learning is promising.

A final remark pertains to the issue of explanatory levels. As before, we did not discuss the level of description of variables that show up in causal networks. This is irrelevant to the methods on offer: all variables, from biological and behavioral to cognitive and social, are treated on a par by the causal modeling methods. Much like model selection methods, causal networks therefore offer a grip on multi-level disease classification.

5. A-reductionism in psychiatry

We are now in a position to relate the current chapter to the central theme of this book on explanatory levels. Psychiatry is inherently multilevel; risk factors for psychiatric disorders are widely dispersed across biological, psychological, and environmental levels (Kendler 2013, Kendler 2014). Over two-thirds of studies have a within-level focus (Kendler 2014), and some researchers give a higher priority to factors from one level, such as genetic factors or other

biological factors (cf. Eronen 2019). But the levels all have their own concepts and their own relevance to the core issue of psychiatry. The benefit of our data-driven approach to classification is that we can involve multiple levels in nosological reform. We do not judge any level as *a priori* more important for classification.

The idea that science needs to be done in terms of concepts stemming from one particular domain or level is very influential. Most commonly this is the domain of the physical, e.g., of the cells and their composition. An important motivation sometimes given for this is metaphysical: science should only be concerned with what exists, and one particular level, typically the material one, has a unique claim to existence. By contrast, epistemic reductionism is the claim that there is a single domain, again typically the material, in terms of which we can ultimately predict, control and explain all the facts about the target system, in our case psychiatric disorders. Weaker versions of the same idea might admit that concepts from other domains will come in handy when explaining the facts, but minimally they will maintain that these concepts are explanatory in virtue of a translation that can be made towards the concepts belonging to the fundamental domain.

So-called epistemic anti-reductionism denies that one single domain takes explanatory priority. The positive claim of the anti-reductionist is that for a full understanding of psychiatric disorders we need concepts from different domains. An important argument for this view is that facts expressed in terms of different domain vocabularies require explanations in terms of those different domains, because the domain vocabularies contain terms that resist translation. This radical untranslatability even motivates some theorists to endorse metaphysical anti-reductionism. If some facts can only be explained by reference to, e.g., concepts from the social

domain, then it might seem reasonable to bestow some kind of existence upon the referents of those concepts as well.

The pragmatic and empiricist position that we have developed in the foregoing is deliberately non-committal when it comes to the debate over epistemic reductionism: it is an a-reductionist point of view. The positive claim we want to make, is that this reductionist issue can be resolved empirically and pragmatically. In search of a better classification of mental disorders, we can take all manner of variables into consideration. The methods that we have advertised in order to decide over inclusion into the classification scheme do not favor one domain over another, and they do not restrict us to a specific domain at the outset. Relative to the predictive and interventionist goals that we set ourselves, we will find that some set of variables will perform best. For certain disorders this may turn out to be a set of variables stemming from one single explanatory domain, whereas for other disorders it will be a set that includes variables from multiple domains. That will all depend on which variables will benefit the goals of the classification scheme, viz. prediction and intervention, most.

Perhaps this attitude seems to preselect a classification scheme that includes variables from multiple domains. Do we not have any principled reasons to prefer a classification scheme that involves only one such domain? Clearly, if we think that certain concepts help science progress faster, take precedence metaphysically, or fit better with our overall world view, then the choice of variables should be constrained. But our point is precisely that adopting such constraints at the outset, by choosing a set of natural kinds for instance, is unnecessary, and that

it might hamper our ability to predict and intervene.[8] If we find use for intervention variables on the explanatory level of behavior, or if it turns out that certain biological, social or cultural factors help us to predict the course of illness, then we should avail ourselves of these conceptual means, without regard for the explanatory level from which they originate.

The drive towards a physicalist vocabulary in psychiatry will in some cases be motivated by a sentiment of "smallism" and "physics envy". This refers to the idea that descriptions in terms of component parts are always more fundamental, better for scientific progress, or that the components have more of a claim to reality than the composites.[9] It is a view often associated with the natural sciences, in particular with fundamental physics. However, as the history of the natural sciences abundantly shows, the search for adequate concepts is ultimately an empirical matter (e.g., Kuhn 1962), and the design and selection of these concepts is arguably what made these sciences so successful. It would be an error to rob psychiatry of one of science's most effective means to support prediction and intervention, namely the freedom to come up with new conceptions of classification criteria. The characteristics eligible for inclusion in psychiatric classification range from bio-markers to environmental factors, and a drive towards the micro-level will only stand in the way of making optimal classification choices.

6. Conclusion

[8] Tabb (2017) argues that the centrality of the DSM presents obstacles to the identification of salient patient groups, and Tabb ([this volume]) presents arguments against a particular take on precision-medicine. Our views are similar to hers in that she argues against taking one specific conceptual schema as the be-all and end-all of psychiatric science. Our attitude here should be a pragmatic one.

[9] Turkheimer [this volume] also critically discusses a tendency to take the smaller scale as more fundamental, arguing that we should instead search for the level of description that brings our entities sharply into focus. We agree with this but replace entities by statistical relations.

The foregoing was informed by a discussion from the philosophy of science which, we argue, sits at the very centre of all classification efforts. It is the discussion on the so-called reference class problem, the problem that the ascription of chances to an individual requires us to see the individual as a member of a group of similar individuals. In the case of psychiatry, as explained in section 2, the reference class problem is that we can only determine the chance of recovery of an individual patient once we have located the patient in a group of relevantly similar patients. Classification schemes aim to provide us with such homogeneous patient groupings.

The identification of the problem of psychiatric nosology with the reference class problem suggests a specific approach to nosological reform. Broadly speaking, if the problem of classification is one of finding statistically homogeneous patient groups, specific statistical methods may help us to identify such groups. Sections 3 and 4 discussed two such statistical methods, to wit, the construction and selection of statistical models and the statistical analysis of causal relations respectively. While these discussions will not build a complete case for the application of these methods and remain rather programmatic, we hope that they will invite researchers to frame their research efforts in the way that we outline, and reconsider the statistical methods that serve their goals. We do not claim warranted optimism about finding intra-homogeneous and inter-heterogenenous patient groups in this way, but we think our approach to be a plausible way forward.

Another benefit of these methods is that they can deal with the inherent multilevel nature of risk factors that are implicated in psychiatry, such as biological, psychological, and environmental risk factors. We support an empiricist and pragmatic approach to psychiatric nosology, in which inclusion of a characteristic into a classification scheme depends on whether

or not this improves our predictions or interventions. When something is an improvement may vary from one classification effort to another, but the goal of prediction and intervention is generic. Most relevant to the theme of this book, there is no presupposition on the so-called explanatory level that the characteristic is associated with. The classification scheme may include characteristics from a multitude of levels if this is what serves the purpose of predicting and intervening best.

Acknowledgements

We would like to thank Ken Kendler, Kathryn Tabb, Eric Turkheimer and Peter Zachar for helpful discussions.

REFERENCES

American Psychiatric Association (2013) *Diagnostic and statistical manual of mental disorders, fifth edition: DSM-5*. 5th edn. American Psychiatric Publishing: Washington, DC.

Baumeister H, Parker G (2012) 'Meta-review of depressive subtyping models.' *Journal of affective disorders* 139, 126–140.

Cartwright N, Hardie J (2012) *Evidence-based Policy*. Oxford University Press.

Chalupka K, Eberhardt F, Perona P (2017) 'Causal feature learning: an overview.' *Behaviormetrika* 44:137–164.

Claeskens G, Hjort N (2008) *Model Selection and Model Averaging.* Cambridge University Press.

Hájek A (2007) 'The Reference Class Problem is Your Problem Too.' *Synthese* 156: 185-215.

Hathaway SR, McKinley JC (1940) 'A multiphasic personality schedule (Minnesota): Construction of the schedule.' *Journal of Psychology*, 10, 249-254.

Eaton WW, Shao H, Nestadt G, Lee BH, Bienvenu OJ, Zandi P (2008) 'Population-based study of first onset and chronicity in major depressive disorder.' *Archives of General Psychiatry* 65, 513–520.

Eronen MI (2019) 'The levels problem in psychopathology.' *Psychological Medicine*, in press.

First MB, Kendler KS, Leibenluft E (2017) 'The Future of the DSM Implementing a Continuous Improvement Model.' *JAMA Psychiatry* 74, 115.

Glymour C, Spirtes P, Scheines R (2001) *Causation, Prediction, and Search (2nd edition).* MIT press.

James G, Witten D, Hastie T, Tibshirani R (2013) *An Introduction to Statistical Learning with Applications in R.* Springer: New York.

Kendler KS (2013) ‘A history of the DSM-5 scientific review committee.’ *Psychological Medicine* 43, 1793–1800.

Kendler KS (2014) ‘The structure of psychiatric science.’ *The American Journal of Psychiatry* 171, 931–938.

Kessler RC, van Loo HM, Wardenaar KJ, Bossarte RM, Brenner LA, Cai T, Ebert DD, Hwang I, Li J, de Jonge P, Nierenberg AA, Petukhova M V, Rosellini AJ, Sampson NA, Schoevers RA, Wilcox MA, Zaslavsky AM (2016) ‘Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports.’ *Molecular Psychiatry* 21, 1366–1371.

Kessler RC, Ustun TB (eds) (2008) *The WHO World Mental Health Surveys: Global Perspectives on the Epidemiology of Mental Disorders*. Cambrdige University Press: New York, N.Y.

Kuhn, T. (1962) *The structure of Scientific Revolutions*. Chicago University Press.

Kupfer DJ, Regier DA, Kuhl EA (2008) ‘On the road to DSM-V and ICD-11.’ *European Archives of Psychiatry and Clinical Neuroscience* 258, 2–6.

van Loo HM, Cai T, Gruber MJ, Li J, de Jonge P, Petukhova M, Rose S, Sampson NA, Schoevers RA, Wardenaar KJ, Wilcox MA, Al-Hamzawi AO, Andrade LH, Bromet EJ, Bunting B, Fayyad J, Florescu SE, Gureje O, Hu C, Huang Y, Levinson D, Medina-Mora ME, Nakane Y, Posada-Villa J, Scott KM, Xavier M, Zarkov Z, Kessler RC (2014) ‘Major depressive disorder subtypes to predict long-term course.’ *Depression and anxiety* 31, 765–777.

van Loo HM, Cai T, Gruber MJ, Li J, De Jonge P, Petukhova M, Rose S, Sampson NA, Schoevers RA, Wardenaar KJ, Wilcox MA, Al-Hamzawi AO, Andrade LH, Bromet EJ,

Bunting B, Fayyad J, Florescu SE, Gureje O, Hu C, Huang Y, Levinson D, Medina-Mora ME, Nakane Y, Posada-Villa J, Scott KM, Xavier M, Zarkov Z, Kessler RC (2014) 'Major depressive disorder subtypes to predict long-term course.' *Depression and Anxiety* 31, 765–777.

Lubke GH, Muthén B (2005) 'Investigating population heterogeneity with factor mixture models.' *Psychological Methods* 10(1): 21-39.

Grove WM, Meehl PE (1996) 'Comparative Efficiency of Informal (Subjective, Impressionistic) and Formal (Mechanical, Algorithmic) Prediction Procedures: The Clinical–Statistical Controversy.' *Psychology, Public Policy, and Law* 2: 293–323.

Morgan M, Morrison M (1999) Models as Mediators. Cambridge University Press.

Nandi A, Beard JR, Galea S (2009) 'Epidemiologic heterogeneity of common mood and anxiety disorders over the lifecourse in the general population: a systematic review.' *BMC Psychiatry* 9, 31.

Olbert, C.M., G.J. Gala, L.A. Tupler (2014) 'Quantifying Heterogeneity Attributable to Polythetic Diagnostic Criteria: Theoretical Framework and Empirical Application.' *Journal of Abnormal Psychology* 123:2, 452–462

Pearl J (2000) *Causality*. MIT press.

Pearl J (2018) *The Book of Why*. New York: Basic Books.

Reichenbach, H. (1949) *The Theory of Probability*. University of Chicago Press.

Roy-Byrne PP, Stang P, Wittchen H-U, Ustun B, Walters EE, Kessler RC (2000) 'Lifetime panic–depression comorbidity in the National Comorbidity Survey.' *British Journal of Psychiatry* 176, 229–235.

Tabb, K. (2015) ‘Psychiatric Progress and the Assumption of Diagnostic Discrimination.’ *Philosophy of Science* 82: 1047–1058.

Tabb, K. (2015) ‘Philosophy of psychiatry after diagnostic kinds’ *Synthese*. DOI: 10.1007/s11229-017-1659-6

Wardenaar KJ, van Loo HM, Cai T, Fava M, Gruber MJ, Li J, de Jonge P, Nierenberg AA, Pethukova M V, Rose S, Sampson NA, Schoevers RA, Wilcox MA, Alonso J, Bromet EJ, Bunting B, Florescu SE, Fukao A, Gureje O, Hu C, Huang YQ, Karam AN, Levinson D, Medina-Mora ME, Posada-Villa J, Scott KM, Taib NI, Viana MC, Xavier M, Zarkov Z, Kessler RC (2014) ‘The effects of co-morbidity in defining major depression subtypes associated with long-term course and severity.’ *Psychological Medicine* 44, 3289–3302.