Evaluating psychodiagnostic decisions

Cilia L. M. Witteman MPhil PhD,1 Clare Harries PhD CPsychol,2 Hilary L. Bekker MSc PhD CHealthPsychol3 and Edward J. M. Van Aarle MSc PhD CPedagogue4

1Professor of Psychodiagnostic Decision Making, 2Lecturer in Psychodiagnosis, Behavioural Science Institute, Radboud University, Nijmegen, the Netherlands
3Lecturer in Psychology, Department of Psychology, University College London, London, UK
4Senior Lecturer in Behavioural Sciences, Academic Unit of Psychiatry and Behavioural Sciences, University of Leeds, Leeds, UK

Introduction

When Di Caccavo & Reid [1] asked 36 General Practitioners how they made patient management decisions, five of them indicated that they did not have access to their own decision processes, stating for example ‘...it all goes into a dark box and makes a decision – which sometimes mystifies me...’ Eleven indicated the same but more subtly: they explained how they negotiated management decisions to their patients, once they themselves had already decided on courses of action. Twelve gave general explanations, talking in terms of tailoring the decision to the patients’ needs and expectations or using their medical training and experience. Eight of the doctors gave scientific theories about their decision making, including references to pattern recognition, algorithms, decision trees and personal protocols.

Just as these General Practitioners did, researchers in judgement, decision making and reasoning, coming from different disciplines within and outside psychology have used several theoretical frameworks to describe and evaluate clinicians’ cognitive behaviour [2]. For example, diagnostic decision making has been framed as using exemplars, prototypes or forceful-feature driven pattern recognition (e.g. Genero & Cantor [3]; Elstein [4]); as forward or backward logical reasoning (e.g. Patel & Groen [5]), as hypothesis testing (e.g. Elstein et al. [6,7]) and as statistical reasoning (e.g. Eva & Brooks [8]); as information integration and use (e.g. Harries et al. [9]; Dhami & Harries [10]); as reason-based [11]; as story [12,13] or script based [14–16]; as based on attitudes and beliefs [17]; and as based on the use of heuristics [18].

From one perspective such theoretical frameworks coexist because they are simply different rhetorical tools: paramorphic descriptions of the same underlying process, as are all models [19]. The same clinician’s behaviour might be framed as pattern recognition, as story-telling, or as combining information, depending upon the particular interests of the researcher and the researcher’s audience. From another perspective, different descriptions may refer to the behaviours of clinicians of different levels of experience, who are experiencing different task conditions, or who perform different sub-tasks of the diagnostic process [20,21]. For example, Einhorn et al. suggest that a think aloud...
description emphasizes the attention to and selection and interpretation of information, whereas an analysis based on a regression model captures its use and integration [22]. And Juslin et al. demonstrated that categorical judgements (such as classifying a problem as anorexia) are best described as exemplar-based pattern matching, while for judgements of a probabilistic continuous criterion (such as a judgement of the risk of self harm) the best description is in terms of cue integration [23].

In this paper we discuss the theoretical frameworks that have been and can be used to evaluate diagnostic decision making.

Evaluating psychodiagnostic decision making

The psychodiagnostic process has three parts: categorical diagnosis, explanatory diagnosis and treatment decision. Evaluating the whole process at once is difficult, as it consists of many different sub-tasks. Most researchers have evaluated the first part of the process: they have evaluated psychodiagnosis in terms of the classification of a client’s problems in for example a Diagnostic and Statistical Manual (DSM) category [24]. Treatment decisions can be evaluated with effect studies, which are carried out by another group of researchers who are dedicated to improving practice. We do not address treatment decisions in this contribution. Rather, in the next sections we first discuss the evaluation of diagnostic categorization and then that of diagnostic explanation. Clinicians’ explanatory diagnostic behaviour is insufficiently understood [25,26] and as far as we are aware it has never been evaluated systematically.

Evaluating categorization performance

Categorical psychodiagnosis can be evaluated both in terms of the match to a criterion classification (an outcome measure), or in terms of the process by which the categorization is assessed. In terms of the outcome of the categorical diagnosis, evidence-based knowledge is even harder to obtain in the domain of mental health than in that of physical health, which results in low expert consensus. In terms of method, clinicians have been made to look bias-prone and erratic, from an information integration tradition, a hypothesis testing and a Bayesian tradition. The gold standard for psychodiagnostic classification, the DSM, is explicitly a-theoretical. But that does not make it an easy standard to live up to: the DSM category can not simply be read off the client. It needs to be assessed by a clinician, and a plethora of research on clinicians’ assessment suggests that their performance on this task is suboptimal (for reviews see Dawes [27]). Almost without exception, empirical studies show that different clinicians are incapable of providing equivalent classifications for the same diagnostic problem, and inter-rater reliability and test–retest consistency (both silver standards of expertise [28]) is limited [29]. This inconsistency in judgement has been credited as the cause of the renowned bootstrapping effect, whereby linear models of clinicians’ judgements consistently outperform the clinicians’ own judgements (cf. Meehl [30]; Grove et al. [31]). In addition, research suggests that there is little difference between the psychodiagnostic classifications of newly qualified psychology students (who have never seen a client) and expert clinical psychologists [32]. Although there are individual differences (e.g. Spengler & Strohmer [33]), in general performance is poor. This can be seen as a consequence either of deficiencies in cognitive processing on the part of the clinicians, or of the inherent difficulty of a classification task in which the environmental information about clients or the pertaining scientifically established knowledge is noisy, gold standards are not agreed, feedback is hard to come by and decision aids are sparse if available at all [34].

The cognitive processes by which clinicians reach their categorizations have been studied from different approaches. They all describe the process by comparing it to a norm: statistical models, hypothetico-deductive reasoning, or Bayes’ theorem.

Statistical models

Early studies of psychodiagnostic classification used a comparison to linear models for weighting and combining information. This has resulted in recognition of the ‘robust beauty’ of linear models [35], against which clinical judgement appears suboptimal. It also allows a comparison of the information used by the judge (that which is significant in the model of their behaviour) both with the information listed in the DSM, and with a model of other judges. Researchers have examined what information judges use, and how they weight and integrate it when making a decision [36]. In such studies, clinicians typically make classification decisions for a series of cases, and most often regression analyses are used to model their decision behaviour. This yields a description of a clinician’s decision in terms of the cues and the cue weights that are statistically significant predictors of the decision, with a function that relates the cues to the decision [10]. In a classic study, Meehl [37] (see also Goldberg [32]) collected data from 29 clinical psychologists and clinical psychology students who were asked to judge 861 psychiatric clients as neurotic or psychotic from their clinical profiles. All judges were presented with the same information, but different statistical models were most accurate in predicting the judgements of different clinicians. This means that the cues they picked up and used, also differed correspondingly.

Hypothetico-deductive reasoning

Another tradition is to compare clinicians’ diagnostic classification processes with a normative model of hypothetico-deductive reasoning. Generally, studies in this tradition focus on which reasoning heuristics diagnosticians use, and which biases in their decision making these bring about (for a comprehensive overview see Garb [26]). Many experiments reveal that clinicians’ decisions are affected by the presentation of information. To give one of many possible examples: Garb demonstrated that the representativeness heuristic described how clinicians make diagnoses [38]. His participants compared a client to a typical client, whose symptoms were representative of those of a specific disorder (e.g. schizophrenia). The presented information is more readily accessible than more valid (experiential or textbook) clinical information, and this heuristic information is substituted for the more valid information when generating a preliminary categorization (cf. Kahneman [39]). This classification seems plausible to the clinician given the information used, and clinicians are reluctant to adjust away from it (see also Friedlander & Stockman [40]).
Bayesian reasoning

Diagnosticians’ categorizing performance has also often been checked against another normative model: Bayes’ theorem. Bayes’ theorem should be used to calculate the probability of alternative classifications given a set of symptoms and test results, and to update the probability of classifications given new information. From this Bayesian perspective, judgement of the likelihood of a particular classification should be based on diagnostic information, that is, information that differentiates between diagnoses (for example: hallucinations may occur with a psychosis but not with a depression), rather than on typical information, that is, information that is typical for one diagnosis but that does not differentiate that diagnosis from others (for example, depressed clients typically sleep badly, but so do traumatized clients). When people focus on typical information they may lose sight of relevant differences, and hence they may become less effective in classifying diseases correctly. Jansen et al. demonstrated that clinicians perform such pseudo-diagnostic reasoning [41]. These clinicians focused on typical instead of diagnostic information, in a fairly complex diagnostic decision-making task. This un-Bayesian behaviour is encouraged by the current format of information in clinical textbooks: diseases are characterized separately in terms of their typical symptoms.

Conclusions on categorical psychodiagnosis

We may conclude, from the above, that psychodiagnosticians’ unaided categorizing behaviour is not coherent with normative models. This should not be surprising, when one realizes which assumptions underlie these models. The use of regression models presumes that information is weighted linearly and combined additively. This may not be a realistic description, although nonlinear versions of regression models have added little predictive power [42]. Both linear and configural regression models assume that a clinician uses the same weighting and combination rule on each case. This seems at best implausible. Recently, fast and frugal models, that is, simple process models that base their decision on only one cue, have successfully been used to describe clinicians’ behaviour [10,43]. These models do not assume that the same information is weighted and integrated on every case (except when there is only one cue in the model), and they incorporate stopping rules, such as to stop when the first cue that differentiates between alternatives is found. They often predict behaviour as well as linear models do, and may be more realistic (though see Harries & Dhami [44]).

Clinicians have also been found to be deficient hypothesis-testers and faulty Bayesians [26,27,45], depending on the cognitive sub-task studied. These two models assume that people have unlimited working memory space and cognitive capacity, while it is by now an established fact that they are boundedly rational, and do not optimize but satisfice [46]. Indeed, this suboptimality in clinicians’ performance has been recognized, and a task force has been set up that has given prescriptions for psychodiagnosticians how to perform systematic hypothesis-testing [47]. Indeed, the studies evaluating diagnosticians’ categorizing performance illustrated above make it clear that debiasing the clinicians and/or the environment is called for.

Describing explanatory diagnosis

An explanatory diagnosis consists of the hypothesized causes of the client’s problems. A diagnostician will for example assume that a traumatic experience explains a phobia. Such an explanatory diagnosis forms a separate part of the process, for example as a means of communicating with the client or colleagues, or it could inform the treatment decision. In the treatment of the phobia, the preceding traumatic experience may be discussed, and a treatment of a disorder classified as attention deficit hyperactivity disorder will be different depending on whether brain damage or poor attachment is assumed to be its cause. Causal explanations may also influence the diagnostic category. For example, Kim & Ahn found that expert clinical psychologists’ categorization was based on their idiosyncratic causal theories. A good causal explanation will always refer to a psychological theory that supports the assumed causal relation [48].

An explanatory diagnosis is a causal model of the client’s problem, or, in terms used by naturalistic decision scientists, a situation assessment. Klein et al. describe how decision makers assess the situation, how they create a mental model of the data and how they then, using their knowledge and experience and with this situation assessment in mind, make the most plausible decision [49,50].

Situation assessments have also been studied in the context of jury decision making (e.g. Pennington & Hastie [51]; Hastie & Pennington [52]). These researchers propose a ‘story model’, in which situation assessments are understood to take the form of a coherent story. Coherence combines completeness, consistency and plausibility. All elements in the story are causally and temporally connected to the central question, for example, ‘what is the matter with this patient?’ and there are no loose or contradictory elements.

This story model may easily be transposed to diagnostic decisions, which share with legal decisions that they both concern a multitude of heterogeneous information about a person’s behaviour, motivations and intentions. In psychodiagnosis, a story, just as other formats of diagnostic explanation, requires reference to a theory or model with causal rules. A good story is a narrative that explains the complaint in terms of temporal and causal arrangements of events in the client’s life [53]. No story should be accepted only on the basis of its plausibility. Apart from being founded in theory, a story should be coherent and consistent. Such measures are each a matter of subjective judgement. Elements of the stories may subsequently be treated as hypotheses, and processed in a hypothetico-deductive manner. If for example the story contains the hypothesis that maladaptive behaviour is caused in part by language difficulties, a test for language comprehension may be performed. Thus, the story may indicate which tests to order, and clients are observed and interviewed, to adapt and refine the tale. Stories only differ from other causal explanations in their reference to a temporal order of events. This may or may not be important for the subsequent treatment decision, but communication is easier in terms of stories than in terms of cause–effect links.

In the medical domain, expert clinicians are found not to construct causal explanations but to recognize scripts: pre-stored, generic causal knowledge structures, built up with experience [14–16]. Medical scripts are schemas associated with causal sequences of events that occur frequently in a specific order in an illness, and
they are activated automatically upon seeing a client, through an unconscious process of memory association. Scripts are high-level, skeleton causal models. Just as medical doctors do, psychodiagnosticians may also recognize scripts: compilations of their theoretical clinical knowledge and prior experience in skeleton stories. As psychodiagnosticians have to explain their decision to clients and colleagues, they will fill in such a script with particulars of the client under consideration, to tell the full individual story. A story, script-based or idiosyncratic, gives coherence to the presented information, and allows explanation of what went on before and prediction of what will follow [54].

**Evaluating causal explanations**

As described above, diagnostic explanations consist of causal propositions, possibly ordered in a story or script. In theory, the basic facts about the client contained in the causal explanation could be checked against reality. In fact, a client’s reality is obviously subjective, which makes such a check unrealistic. Another option is to assess the format of the explanations: do they adhere to certain formatting requirements? Researchers in several disciplines (statistics, logic and philosophy) have suggested ways in which this might be done.

**Statistics**

Causal explanations can be evaluated by representing them in a qualitative or quantitative probabilistic network [55,56] or influence diagram: a directed probabilistic network with nodes for (client) data, decision nodes and causal relations connecting all nodes that influence each other. The impact of changes in decision node is easily calculated, giving statistically correct results. An influence diagram may be used as an expert model of the causal explanation (cf. Morgan et al. [57]).

**Logic**

Causal explanations can also be evaluated in terms of their logic, a technique often used in computer science and artificial intelligence. There are four plausible methods with which to perform this evaluation. The first is to apply a Logic of Argumentation (e.g. Parsons & Fox [58]) which tests the logical correctness of conclusions given their arguments. The second is to analyse the process as a dialogue game [59] between the clinician and the client and/or other informants. The third is analysing decision protocols with a story grammar [60] which proposes a set of syntactical rules and a corresponding set of semantic interpretation rules to determine the ‘well-formedness’ of the explanation; and the fourth is to apply Toulmin’s Theory of Logic [61], which makes the logic underlying an argument explicit and by which one is able to state whether the conclusion is backed by evidence.

**Philosophy**

In philosophy, the test for the validity of a causal explanation would be to reason counterfactually. Grieving for a lost loved one is caused by the actual loss of that loved one. A psychodiagnostician has to be able to justify, referring to a psychological theory, that without the previous event, the behaviour could not have occurred. In fact, in psychodiagnostics, one should allow for multiple causes. For many behaviours it is hardly imaginable that they would be the result of one isolated event only. A depression for example may be caused by loss of a loved one and a passive coping style and an otherwise bleak personal situation. A cause may be a condition with elements that together necessitate the effect.

**Conclusions on causal explanation**

A number of – as yet untried – methods is available to evaluate psychodiagnostic explanations by looking at their structure, not their easily contestable content. This avoids disputes about what is real or objectively true about a client’s problems. We argue that application of such methods deserves investment of effort by researchers.

**General conclusions**

In this essay we have described different frameworks to evaluate psychodiagnostic decision making, which address different aspects of the decision. Evaluation so far has focused on the first part of the process: categorial classification, using statistical models, hypothetico-deductive models and Bayes’ theorem. The resulting assessments with any of these models have not been complimentary. We have argued that the models make unrealistic assumptions about the human decision makers. Moreover, a conclusion that psychodiagnosticians are poor decision makers may be premature, because to the best of our knowledge, it has never been established whether following linear models, or Bayes’ theorem, or hypothetico-deductive prescriptions leads to better diagnoses. More importantly, evaluation studies have not addressed causal explanations. We argue that this is an important area of research that is still quite blank. We have sketched a number of models that may be applied in this area.

Once both these parts of the psychodiagnostic process have started to be studied together, the effectiveness of different normative methods to aid improvement can be assessed. The bottom line is: how are the clients best served? If effect studies show differences in therapeutic gain when clinicians follow different methods of classification and explanation, then the best performing methods should be taught. As no such effects have yet been established, psychodiagnosticians and clients are well served with further descriptive and evaluative research.

**References**

Evaluating psychodiagnostic decision making


