Psychological Medicine (2016), **46**, 2255–2262. © Cambridge University Press 2016
doi:10.1017/S003329171600074X

REVIEW ARTICLE

Expert consensus *v.* evidence-based approaches in the revision of the DSM

K. S. Kendler[1,2,3]* and M. Solomon[4]

[1] *Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, VA, USA*
[2] *Department of Psychiatry, Virginia Commonwealth University, Richmond VA, USA*
[3] *Department of Human and Molecular Genetics, Virginia Commonwealth University, Richmond, VA, USA*
[4] *Department of Philosophy, Temple University, Philadelphia, PA, USA*

The development of DSM-III through DSM-5 has relied heavily on expert consensus. In this essay, we provide an historical and critical perspective on this process. Over the last 40 years, medicine has struggled to find appropriate methods for summarizing research results and making clinical recommendations. When such recommendations are issued by authorized organizations, they can have widespread influence (i.e. DSM-III and its successors). In the 1970s, expert consensus conferences, led by the NIH, reviewed research about controversial medical issues and successfully disseminated results. However, these consensus conferences struggled with aggregating the complex available evidence. In the 1990s, the rise of evidence-based medicine cast doubt on the reliability of expert consensus. Since then, medicine has increasingly relied on systematic reviews, as developed by the evidence-based medicine movement, and advocated for their early incorporation in expert consensus efforts. With the partial exception of DSM-IV, such systematic evidence-based reviews have not been consistently integrated into the development of the DSMs, leaving their development out of step with the larger medical field. Like the recommendations made for the NIH consensus conferences, we argue that the DSM process should be modified to require systematic evidence-based reviews *before* Work Groups make their assessments. Our suggestions – which would require leadership and additional resources to set standards for appropriate evidence hierarchies, carry out systematic reviews, and upgrade the group process – should improve the objectivity of the DSM, increase the validity of its results, and improve the reception of any changes in nosology.

Received 7 December 2015; Revised 15 March 2016; Accepted 15 March 2016; First published online 13 April 2016

Key words: DSM, evidence-based medicine, expert consensus.

Introduction

The Diagnostic and Statistical Manual of the American Psychiatric Association (DSM) is an evolving document that seeks to classify psychiatric disorders in ways that satisfy both research and clinical needs. This paper looks at how revisions to the DSM have been made from DSM-III (APA, 1980) onwards, and makes recommendations for improving this process. The issues are viewed in the context of more general developments in the making of medical knowledge over the last 40 years, in particular the consensus conference movement and the rise of evidence-based medicine (Solomon, 2015).

From DSM-III onward, the primary deliberative bodies for developing recommendations for change have been groups of individuals (ranging typically from four to around 15) chosen for their clinical and research expertise in specific diagnostic areas. These bodies, called 'Advisory Committees' in DSM-III (APA, 1980) and DSM-III-R (APA, 1987), and 'Work Groups' in DSM-IV (APA, 1994) and DSM-5 (APA, 2013), were asked by the DSM leadership to review research and clinical practice in the relevant diagnostic areas and make suggestions for possible changes. These suggestions were then discussed among members of the Work Groups and the pros and cons of making particular changes weighed. Through discussion and an informal consensus process, each Work Group produced reports with recommendations for change. The reports were forwarded to the oversight committee – typically called the 'Task Force' – whose members comprised DSM leadership and Chairs of the various Work Groups. From there, approved proposals for change were put to the APA Board of Trustees and Assembly for final approval. The DSM is thus the product of an ordered sequence of group discussions and deliberations to consensus.

A variety of kinds of information were reviewed in the Work Group discussions. Considerations were conceptual, empirical, and/or clinical. Conceptual considerations might include arguments that a particular way of constructing a category was unclear or inaccurate and could be improved upon. Empirical

* Address for correspondence: Professor K. S. Kendler, VCU, Box 980126, Richmond, VA 23298-0126, USA.
(Email: Kenneth.Kendler@vcuhealth.org)

considerations might include arguments that the criteria for a disorder should be changed in order to improve validity or reliability. Other times clinical issues were discussed such as certain criteria being too complex for easy clinical use. Sometimes, new research findings that addressed the validity of specific proposed changes were introduced. In some Work Groups, thorough literature reviews were done during the deliberative process so that they could influence the proceedings.

Surprisingly, literature reviews were more systematically encouraged for DSM-IV than for DSM-5; DSM-IV even had a Methods Conference to establish guidelines for Task Force members to do these reviews, and planned external review for all the literature reviews (Widiger *et al.* 1990). DSM-5 was less explicit about the need for literature reviews as a basis for Work Group proposals for change.

This paper will argue that while the reliance on expert consensus might have been appropriate when DSM-III was first published in 1980 (APA, 1980), it should now be reduced in favor of more systematic evidence-based approaches. The leadership of DSM-IV presciently called for this in 1990 (Widiger *et al.* 1990) – before the term 'evidence-based medicine' was introduced in 1992 (Evidence-based Medicine Working Group, 1992) – but did not fully implement it, and the methods of systematic evidence review have developed considerably since then, requiring updating of the steps to take in doing a systematic evidence review. Expert consensus still has an important role to play in making clinical and policy recommendations, and should be improved by incorporating insights from the relevant literature on processes of group deliberation. We give an overview of the development of both medical consensus conferences and evidence-based medicine, and discuss the implications for the DSM process.

Authority and expertise

Reliance on expertise is an ancient and universal method for authenticating medical knowledge. From the time of Greek medicine until the Scientific Revolution, invoking the writings of Hippocrates and Galen along with teaching in a master–apprentice relationship – which both involve appeals to expert authority – were the dominant ways of producing and communicating medical knowledge. In this historical context, reliance on one or two giants (Hippocrates and Galen, in the case of medicine) was typical.

Expert consensus is a more contemporary kind of reliance on authority. It involves reliance on a group of experts who are, or come to be, in agreement with each other. It considers that the agreement of experts is an additional reason for confidence in the reliability of their joint agreement – if all the experts agree, who are we to disagree? Expert consensus is a combination of earned authority (from demonstrated expertise) and democracy (in the careful deliberation of the group to consensus). Much committee work – outside as well as inside medicine – is designed to produce such expert consensus on practical topics.

In the years after World War II, medical research was generously funded, particularly at the National Institutes of Health (NIH). As results accumulated, the field of medicine began to struggle with when and how to make recommendations for changes in clinical practice. Such recommendations needed to accomplish two tasks: objective evaluation of the evidence, and effective communication of the recommendations so as to convey legitimacy and authority. At first, it was thought that the two tasks could be addressed in the same way, by expert consensus. In 1977, the first attempt to *formalize* the process of expert consensus in a medical context took place at NIH: the NIH Consensus Development Conference Program.

The NIH Consensus Development Conference Program

The NIH Consensus Development Conference Program began in 1976 with a request to the NIH from Senator Edward Kennedy (then chair of the Senate Subcommittee on Health) to accelerate the transfer of information from NIH-funded research to practicing physicians. During the post-World War II years, NIH funding and research had increased more than tenfold, and there was concern about the uptake of the results. In response to Senator Kennedy's request, the director of NIH, Donald Frederickson, created the 'Office of Medical Applications of Research' (OMAR) which in turn began the Consensus Development Conference Program (Solomon, 2015).

The Consensus Development Conference Process began with the choice of a topic by OMAR. The topics of the first four conferences were 'Breast Cancer Screening', 'Educational Needs of Physicians and Public Regarding Asbestos Exposure', 'Dental Implants: Benefit and Risk', and 'Mass Screening for Colorectal Cancer' (Mullan & Jacoby, 1985). Then, the NIH team would assemble a panel of 10–20 independent clinicians, researchers and research methodologists with expertise in the relevant area. By the term 'independent', it was meant that panelists were not NIH employees, and also that panelists had not yet made a public statement about the issues to be discussed at the conference. In these ways, panelists were expected to be without institutional or intellectual biases (concern about financial biases came later). The meetings

https://doi.org/10.1017/S003329171600074X Published online by Cambridge University Press

occurred over an intense three days. During the first 2 days, panel members listened to and questioned a number of experts on various aspects of the chosen topic (these experts usually had extensive publications in the area). In an executive session on the first night, the panel started work on a draft consensus statement which would then be completed on the second night. The third day would start with a reading and public discussion of the draft consensus which would then be revised by the panel into a final form and presented at a press conference later that day.

The NIH Consensus Development Conference Program was well received from its inception. Those participating in the conferences were typically very positive about the process of producing a consensus document. They felt that expert panels were the best way to assess a body of evidence and come to conclusions about the weight of the evidence, especially when some of it was equivocal. The NIH program was widely adopted and adapted in other national and international contexts, and for the next 20 years, medical consensus conferences were the most widely used means of assessment in medical contexts. In the United States, other medical consensus conference programs developed at the Institute of Medicine, the Blue Cross and Blue Shield Technology Evaluation Center, the US Preventative Service Task Force, and many other organizations.

Work on DSM-III (APA, 1980) was completed in the mid- to late 1970s just at the time at which the medical consensus conference movement was getting started. To our knowledge, there was no formal contact between the NIH consensus conference efforts and the DSM-III (Decker, 2013). Rather, the advisory committee structure grew out of Bob Spitzer's desire to involve a wide range of experts in the DSM-III development and the realization that the most efficient way to organize their efforts would be by broad diagnostic category (APA, 1980; Decker, 2013). By contrast, DSM-II, with which Spitzer had also been involved as a consultant, had only a single committee termed 'The Committee on Nomenclature and Statistics of the American Psychiatric Association' which also presumably worked on a consensus model (APA, 1968).

Concerns about the NIH Consensus Development Conference Program

Although the initial reception of the NIH Consensus Development Conference Program was largely positive, a few people raised concerns about relying on a consensus process for evaluating complex evidence. Ten years after its inception, in a *JAMA* editorial, Itzhak Jacoby, a former acting director of the NIH Program wrote:

> More strict reliance on evidence in consensus development might have been promoted if staff preparation for US conferences routinely included a data synthesis for the panel. On the occasions when such a synthesis was prepared by the staff and accepted by the panel, evidence was well integrated into both the deliberations and the consensus statement … At other times, panels showed unwillingness to rely on the background work. When a data synthesis was unavailable or was not used, and when extensive information was provided by expert presentations at the conference, the difficulty of coping was exacerbated. Probably as a result, some consensus statements show evidence of influence by panelists' assertions of common sense or knowledge of acceptable practices … my observation, as director of the NIH Consensus Development Program from April 1984 through July 1987, of the more than 30 contemporary conferences and review of the recent Canadian experience lead me to place greater value on techniques of data synthesis and to reemphasize commitment to basing consensus strictly on examination of evidence. (Jacoby, 1988, p. 3039)

Jacoby's concerns were prescient, but they were not taken up, and for the next 10 years the NIH program had a period of stability under director John Ferguson. The Institute of Medicine's 1990 study (Institute of Medicine, 1990) of the NIH Consensus Development Conference Program recommended providing panelists with a background report systematically reviewing relevant research. This recommendation was not followed.

One of us (M.S.) has previously summarized these concerns about the NIH Consensus Development Conference program as follows:

> Left alone, panelists were sometimes overwhelmed by the quantity and complexity of information provided, and ended up basing their decision on a few salient studies. The biased nature of these decisions was sometimes noticed and criticized. (Solomon, 2015, p. 47; see also Ahrens, 1985)

A good illustration of the rising concerns about expert consensus is a 1992 paper which compared the recommendations of experts (authors of textbooks and review articles) with already available information from randomized controlled trials (Antman *et al.* 1992). The subject was treatment of myocardial infarction. The authors write:

> Discrepancies were detected between the meta-analytic patterns of effectiveness in the randomized trials and the recommendations of reviewers. Review articles often failed to mention important advances or exhibited delays in recommending effective preventive measures. In some cases, treatments that have no effect on mortality or are potentially harmful continued to be recommended by several clinical experts. (Antman *et al.* 1992, p. 240)

The authors noted a range of factors that predicted the failure of expert opinion to keep pace with the best

https://doi.org/10.1017/S003329171600074X Published online by Cambridge University Press

available empirical data. These included: (i) inadequate attention to the latest results, (ii) inappropriate interpretation of small negative trials, (iii) limited familiarity with particular research methods, and (iv) an overreliance on personal experience especially with over-interpretation of rare events in their own practice.

In 1999, the NIH convened a workgroup chaired by Alan Leshner to evaluate the Consensus Development Conference Program and to provide suggestions for change (Leshner *et al.* 1999). By that time, the evidence-based medicine movement was in full flower. Evidence hierarchies typically rank expert consensus at the bottom of the evidence hierarchy – when they count it at all. Not surprisingly, concerns about aggregating the evidence were raised again, this time in a context that could not be ignored, especially since other medical consensus conference programs (such as those at the Institute of Medicine) had already been modified to include a stage of systematic evidence review. By 1999, the NIH program looked seriously outdated. The Leshner Report was blunt:

> The workgroup felt strongly that the consensus development process itself would benefit from the application of new methods to systematically review data prior to a consensus conference, such as those used in the evidence-based approach to establish practice guidelines in recent years. This technique involves a systematic review of the evidence, and while it would be more time consuming and involve a greater commitment from panel members, would improve the process and the product substantially … . A model such as this would afford the consensus panel a greater opportunity to achieve an evidence-based consensus. Some benefits include the conduct of a formal evidence-based review to inform the process, opportunity for panelists to study the systematic review … and an extended period of time for deliberations and reporting. (Leshner *et al.* 1999, pp. 8–9)

The Leshner Report went on to recommend that the NIH Office of Medical Applications of Research, commission such a report to be studied by panelists in advance. This was implemented by 2001 in a partnership with the Agency for Healthcare Research and Quality (AHRQ). For the remaining years of the NIH Consensus Development Conference Program (until 2013), an evidence report was produced at one of AHRQ's Evidence-Based Practice Centers and shared with the panel about 1 month before the consensus conference, and published along with the results of the consensus conference.

General concerns about consensus conferences

Expert consensus is an appealing process. We would like to believe that a group of experts requested to reach a consensus judgment about a complex matter with multiple and sometimes conflicting sources of information will combine their joint wisdoms and reach the correct decision. However, this may not be what actually happens, especially when the consensus process is informal. While we cannot review thoroughly the large literature on group process and decision making here, we will summarize some results. The best place to start is with the concept of 'Groupthink' as popularized by the social psychologist Irving Janis (Janis, 1983). This is defined as follows:

> Groupthink is a psychological phenomenon that occurs within a group of people, in which the desire for harmony or conformity in the group results in an irrational or dysfunctional decision-making outcome. Group members try to minimize conflict and reach a consensus decision without critical evaluation of alternative viewpoints, by actively suppressing dissenting viewpoints, and by isolating themselves from outside influences. (Wikipedia contributors, 2015)

Janis gives a detailed evaluation of well-known historical examples of errors in decision making (The Bay of Pigs invasion, Pearl Harbor, The Vietnam War; Janis, 1983) which he blames on groupthink. The upshot is that group deliberation sometimes produces worse decisions than can be obtained without deliberation.

Phenomena such as polarization and anchoring also mar group decision making. Polarization (as detected by Myers, 1982) is the tendency for groups to come to conclusions that are more radical than the conclusions that would be reached by individuals alone. Anchoring (as discussed by Tversky & Kahneman, 1982) is the tendency for people to give the first thing said more weight than things of equal validity said later in the discussion, which biases groups towards the opinions of those who speak first.

In addition, attention to diversity of group membership and active cultivation of dissent improve group deliberation. There are other non-intuitive results, such as that non-experts often produce better decisions on a topic than do experts (Surowiecki, 2004), and outcomes of group deliberations are better when the members of groups are strangers rather than colleagues (Sunstein, 2003).

More structured group deliberation, in which steps are taken to avoid phenomena such as groupthink, polarization, and anchoring, can improve the objectivity of the results of group deliberation. Some have attempted to minimize peer pressure by concluding the group discussion with anonymous voting and aggregation rather than consensus building. While this can lead to good outcomes, it loses the practical and rhetorical benefits of consensus.

William Sutherland and Mark Burgman (Sutherland & Burgman, 2015) give a useful up-to-date summary of

https://doi.org/10.1017/S003329171600074X Published online by Cambridge University Press

the problems with relying on expert group processes, and suggest some remedies. These suggestions may be useful when thinking about how best to structure group deliberation in the future, and therefore should be used when devising a process for producing future editions of DSM.

Despite these concerns about the rationality of group process, the NIH program continued with its practices of informal group process. Perhaps the subjective feeling that unstructured discussion is more 'open' than structured discussion played a role in continuing the practice.

Evidence-based medicine

At the time that DSM-III was published, the term 'evidence-based medicine' was not yet in use, and evidence-based reviews did not generally rank evidence in terms of quality (a notable exception was the Canadian Task Force on the Periodic Health Examination, which produced the first evidence hierarchy, in 1979). By the end of the twentieth century, however, evidence-based medicine was widely recognized, and systematic evidence review developed as a set of specialized techniques requiring methodological training as well as significant time and personnel to accomplish.

The term 'evidence-based medicine' is a bit of a misnomer. Medicine has always been (at least in part) based on evidence, broadly defined. What the term really means is 'hierarchy-ranked evidence-based medicine'. Different kinds of evidence, such as randomized controlled trials, observational trials, and case studies, are ranked in terms of quality. In developing recommendations for practice by evidence-based methods, the weight given to particular sources of information is governed by their place in the hierarchy.

Randomized controlled trials, not surprisingly, are at the top. Also, the results of similarly conducted trials can be mathematically combined in a meta-analysis to yield an overall quantitative result. Even when that is not possible, a systematic review of the evidence (complete with literature search) can provide a more objective overview of a body of evidence than 'seat of the pants' judgments. Systematic reviews evaluate both the quality and the quantity of the evidence. Evidence hierarchies are used to help evaluate the quality of the evidence. While they vary to some extent, all rank expert consensus at the lowest level. From the perspective of evidence-based medicine, consensus conferences have little credibility.

An example of an evidence hierarchy is that used by the Oxford Centre for Evidence-based Medicine (2009). It consists of five general levels of evidence and several sub-levels. On level 1 are systematic reviews of randomized controlled trials and individual randomized controlled trials. Level 2 includes systematic reviews of cohort studies and individual cohort studies. Level 3 comprises systematic reviews of case-control studies and individual case-control studies. Level 4 includes case series. Finally, level 5 is composed of 'Expert opinion without explicit critical appraisal, or based on physiology, bench research, or first principles'.

What is interesting for our purposes is that level 5, the bottom category, includes expert consensus. A systematic review, which grades all the evidence and puts together overall recommendations, must regard expert consensus as the weakest form of evidence, equivalent in reliability to theoretical speculations. This is a large change from the years in which the NIH Consensus Development Conference Program represented the cutting edge of assessment of medical knowledge.

Back to the NIH Consensus Development Conference Program

We can now better appreciate the historical context of the concerns that were raised about the NIH Consensus Development Conference Program. The rise of evidence-based medicine in the 1990s with its explicit criticism of the reliability of expert consensus questioned the role of consensus conferences in medicine. The Leshner Report can be understood as trying to move the NIH program away from a pure expert consensus model toward a hybrid between an evidence-based and an expert consensus process. The first step in their proposed model was to present to the expert group, *prior to their deliberations*, all the available data summarized in a manner that would be typical for an evidence-based medicine review. *After having the opportunity to digest this information*, the panel would then hear from experts and develop their consensus. The Leshner panel hoped that this approach would (i) provide the important benefits of an evidence-based medicine approach – a thorough and unbiased review of all available data – and (ii) reduce the problems associated with expert consensus such as selective attention to parts of the literature or over-reliance on clinical experiences. It could here be appropriately asked: why not, then, do away with the expert consensus altogether? Complementary to the more 'technocratic' evidence-based medicine analyses, a review by a distinguished and experienced panel included both an additional level of oversight to the review process and, by adding to the authority and legitimacy of the process, increased the probability of broad acceptance of the final recommendations.

In medical consensus conferences concerned with policy as well as science, expert consensus plays a

https://doi.org/10.1017/S003329171600074X Published online by Cambridge University Press

more determining role, because policy questions cannot be settled by evidence alone. Policy questions also require weighing benefits and harms, considering matters of justice, cost considerations, and so forth. Consensus conferences may be more suited to answering policy questions than to answering scientific questions. An example of a consensus conference program that also considers policy questions is the Medicare Coverage Advisory Committee (since 2007, renamed the Medicare Evidence Development Coverage Advisory Committee); it, for example, considers ethical and economic questions related to the use of medical technologies.

Implications for the DSM

The situation for DSM-5 was similar to that of the NIH Consensus Development Conference Program before the Leshner Report in that there was no *requirement* that DSM Work Groups do (or commission) a systematic evidence review, still less that they did so before meeting as a committee. For DSM-5, some Work Groups produced systematic evidence reviews and some did not. The insertion of the Scientific Review Committee (SRC; Kendler, 2013) after Work Groups was intended as a check on the quality of the scientific support for the changes proposed by the Work Groups. But the SRC was introduced after the Work Groups began their deliberations, and only influenced the results after Work Groups had produced an initial consensus. Ideally, the evidence review should take place *before* the Work Group even meets, so that the discussion is anchored to the evidence rather than to the position of the first person to speak. In addition, the SRC lacked resources to do systematic evidence review, so their check on the science could not be thorough and was limited to a review of the evidence presented to them by the Work Groups and their own limited inquiries.

However, the current situation in the DSM process is different in important ways from that of the NIH Consensus Development Conference Program. First, the evidence base for diagnostic categories in psychiatry is much more disparate than the evidence base for most therapeutic interventions. The evidence base for diagnostic categories includes a range of validators from genetics, imaging, treatment, follow up, and epidemiological studies. Some of these may be randomized controlled trials, but others are not trials at all but rather traditional scientific hypothesis testing studies. As with clinical trials, tests of scientific hypotheses may be of varying quality, but we do not have a clear 'evidence hierarchy' for such studies. In fact, while there is a substantial literature on the use of validators in psychiatric nosology (Kendler, 1990; Robins & Guze, 1970), we do not have explicit standards for how to do a systematic review of the evidence for psychiatric categories. In cases of nosological categories for which we have information on five or six major validators, as well as data on reliability and utility, the results often are not univocal. For example, genetic and neuroimaging findings might support one categorization while treatment and outcome measures support another. We do not now possess an algorithm that can combine these results and reach an undisputed decision on such questions as the relative importance of information on course of illness *v.* treatment, or on how much reduced validity would be tolerable to produce greater reliability or utility. Furthermore, the answers to these questions might vary across different diagnostic categories where the relative importance of false positive *v.* false negative diagnoses may differ. This is an important matter for discussion by those working on future editions of DSM.

Second, the DSM is a clinical as well as a scientific document. In fact, DSM-5 added a Clinical and Public Health Committee (Kendler, 2013) to check Work Group reports for the impact of their recommendations on clinical practice. The DSM process is more like those medical consensus conferences that consider matters of policy as well as matters of science (such as the Medicare Coverage Advisory Committee) than like the NIH Consensus Development Conference Program, which considered solely scientific questions. For such 'non-empirical' matters as the pragmatics of clinical practice, expert consensus is currently the most democratic and effective decision process that we have – especially when the process is structured so as to avoid the sources of bias discussed above.

Recommendations

We recommend that going forward with revisions of DSM, thorough and objective evidence-based reviews play a central part in the process. Furthermore, these reviews should be completed *prior to Work Group deliberations about the import of the evidence* so that they can anchor the discussions and discourage appeal to favorite studies or personal opinions.

Who should complete such reviews? We can imagine three models. Ideally, the reviews required could be identified by Work Groups and then commissioned from an outside impartial agency [e.g. one of the 13 Evidence-Based Practice Centers used by AHRQ (Agency for Healthcare Research and Quality, 2015)]. However, this may not be practical both because of the resources required and the needed specialized expertise. The next best approach would be for the review to be completed by members of the Work Group chosen for their objectivity. That is, they should

https://doi.org/10.1017/S003329171600074X Published online by Cambridge University Press

have a general expertise in the broad issues at hand but would not have strong personal opinions about the specific proposed changes. However, it cannot be assumed that without specific experience or training, a sufficient number of Work Group members will have the skills to do the systematic reviews as the ideal review includes several people so as to reduce individual biases. Staff assistance from the APA might be necessary given the large time commitment involved to complete such reviews. This would take resources but given the importance of the DSM and its widespread sales, it would be hoped that the APA could provide such support. The least desirable option would be to have the reviews completed by the individuals who themselves are making the proposals for change. This obvious disadvantage of this approach is the possibility of bias due to interested reasoning. In our view, this approach is viable only if these reviews are carefully vetted for completeness and objectivity by other individuals not involved in the proposal, perhaps members of the relevant workgroup and/or their staff. The time commitment to vet the reviews would be less than those needed to develop them from scratch, but the confidence in their completeness and lack of bias will be necessarily less.

A further practical problem that will need to be addressed by those designing the next phase of DSM revisions is when a topic is ready for review. Some decision-making system needs to be in place to judge that the proposal is sufficiently meritorious to warrant the work involved. Those who propose the changes can be required to do an initial evidence review to provide prima facie evidence regarding the merit of the proposal. If this is judged credible, then an independent evidence review would be justified.

Thus we are recommending a plan somewhat similar to that proposed for DSM-IV, updated to reflect advances in scientific and meta-analytic methods and to accommodate the specific characteristics of evidence relevant to psychiatric nosology. More attention to the objectivity of group process would also be helpful. Expert consensus is still indispensable for addressing some clinical and policy matters. There are many resources for improving group process, including the summary mentioned above (Sutherland & Burgman, 2015).

Conclusions

In the last 40 years, the field of medicine has struggled with methods for reviewing and summarizing new research results. This is an important issue because of the need in many medical fields for authoritative statements about developments that impinge on the diagnosis and treatment of patients. Indeed, such statements, especially when issued by organizations with recognized authority, can have widespread influence. We certainly see this with the DSM, which has served as a unifying framework for research and clinical practice since the introduction of DSM-III in 1980 (APA, 1980).

We documented two major trends in this 'making of medical knowledge'. In the 1970s, expert consensus, led by the NIH, became an important process for making and disseminating judgments about important medical issues. But by the 1990s, the rise of evidence-based medicine cast doubt on the validity of the expert consensus model. The DSM process, however, did not fully acknowledge this change, and continued to rely largely on the expert consensus model up until and through the recently completed DSM-5 (APA, 2013). Some use of literature reviews has been incorporated but their use is variable, unsystematic, and often up to the discretion of Work Groups.

The wide division that now exists between the general medical community, which has moved the emphasis from expert consensus toward the use of evidence-based methods, and the DSM process should be a source of concern for American psychiatry. Our suggestions for improvement will require leadership and likely additional resources to set evenhanded standards for an appropriate evidence hierarchy(ies) and to carry out the necessary systematic reviews. Moreover, attention to the objectivity of group process will likely improve outcomes. In our judgment, the implementation of such changes would enhance the objectivity of the DSM process, increase the validity of its results, and improve the reception of any changes.

Acknowledgements

Paul Appelbaum MD, Ellen Leibenluft MD, and Jan-Willem Romeijn PhD provided helpful comments on earlier versions of this essay.

Declaration of Interest

None.

References

Agency for Healthcare Research and Quality (2015). Evidence-Based Practice Centers (EPC) Program Overview. August 2015 (internet citation). Agency for Healthcare Research and Quality: Rockville, MD.

Ahrens Jr. EH (1985). The diet-heart question in 1985: has it really been settled? *Lancet* **1**, 1085–1087.

APA (1968). *Diagnostic and Statistical Manual of Mental Disorders*, 2nd edn. American Psychiatric Association: Washington, DC.

https://doi.org/10.1017/S003329171600074X Published online by Cambridge University Press

APA (1980). *Diagnostic and Statistical Manual of Mental Disorders*, 3rd edn. American Psychiatric Association: Washington, DC.

APA (1987). *Diagnostic and Statistical Manual of Mental Disorders – Revised*, 3rd edn. American Psychiatric Association: Washington, DC.

APA (1994). *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. American Psychiatric Association: Washington, DC.

APA (2013). *Diagnostic and Statistical Manual of Mental Disorders*, 5th edn. American Psychiatric Association: Washington, DC.

Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC (1992). A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction. *Journal of the American Medical Association* **268**, 240–248.

Decker HS (2013). *The Making of DSM-III: A Diagnostic Manual's Conquest of American Psychiatry*, 1st edn. Oxford University Press: Oxford, UK.

Evidence-based Medicine Working Group (1992). Evidence-based medicine. A new approach to teaching the practice of medicine. *Journal of the American Medical Association* **268**, 2420–2425.

Institute of Medicine (1990). Institute of Medicine, Consensus Development at NIH: Improving the Program, Report of a Study by a Committee of the Instutite of Medicine Council on Health Care Technology 1990, National Academies Press: Washington, D.C.

Jacoby I (1988). Evidence and Consensus. *Journal of the American Medical Association* **259**, 3039.

Janis IL (1983). *GROUPTHINK: Psychological Studies of Policy Decisions and Fiascoes*, 2nd edn, revised. Houghton Mifflin Company: Boston.

Kendler KS (1990). Toward a scientific psychiatric nosology. Strengths and limitations. *Archives of General Psychiatry* **47**, 969–973.

Kendler KS (2013). A history of the DSM-5 Scientific Review Committee. *Psychological Medicine* **43**, 1793–1800.

Leshner A, Davidson E, Eastman P, Grundy S, Kramer B, McGowan J, Penn A, Strauss S, Woolf S (1999). Report of the working group of the advisory committee to the director to review the Office of Medical Applications of Research, 5 August 2015.

Mullan F, Jacoby I (1985). The town meeting for technology. The maturation of consensus conferences. *Journal of the American Medical Association* **254**, 1068–1072.

Myers DG (1982). Polarizing effects of social interaction. In *Group Decision Making* (ed. H. Brandstätter, J. H. Davis and G. Stocker-Kreichgauer), pp. 125–161, Academic Press: London.

Oxford Centre for Evidence-based Medicine (2009). Levels of evidence, 3 March 2009.

Robins E, Guze SB (1970). Establishment of diagnostic validity in psychiatric illness: its application to schizophrenia. *American Journal of Psychiatry* **126**, 983–987.

Solomon M (2015). *Making Medical Knowledge*, 1st edn. Oxford University Press: New York, NY.

Sunstein CR (2003). *Why Societies Need Dissent*. Harvard University Press: Cambridge, MA.

Surowiecki J (2004). *The Wisdom of Crowds*. Doubleday: New York.

Sutherland WJ, Burgman M (2015). Policy advice: use experts wisely. *Nature* **526**, 317–318.

Tversky A, Kahneman D (1982). Judgment under uncertainty: heuristics and biases. In *Judgment Under Uncertainty: Heuristics and Biases*, 1st edn (ed. D. Kahneman, P. Slovic and A. Tversky), pp. 3–23. Cambridge University Press: Cambridge.

Widiger TA, Frances AJ, Pincus HA, Davis WW (1990). DSM-IV Literature Reviews: Rationale, Process, and Limitations. *Journal of Psychopathology and Behavioral Assessment* **12**, 189–202.

Wikipedia contributors (2015). Groupthink (https://en.wikipedia.org/w/index.php?title=Groupthink&oldid=674546069). Wikipedia, The Free Encyclopedia. 6 August 2015.