

# How common standards can diminish collective intelligence: a computational study

Michael Morreau PhD<sup>1</sup> and Aidan Lyon PhD<sup>2</sup>

<sup>1</sup>Professor of Philosophy, UiT-The Arctic University of Norway, Tromsø, Norway

<sup>2</sup>Assistant Professor of Philosophy, University of Maryland, College Park, MD United States

## Keywords

collective intelligence, computer modelling, evaluation, evidence

## Correspondence

Michael Morreau

Philosophy

UiT Norges arktiske universitet Fakultet for humaniora samfunnsvitenskap og

laerarutdanning

Postboks 6050 Langnes

9037 Tromsø

Norway

E-mail: michael.morreau@uit.no

Accepted for publication: 16 May 2016

doi:10.1111/jep.12585

## Abstract

Making good decisions depends on having accurate information – quickly, and in a form in which it can be readily communicated and acted upon. Two features of medical practice can help: deliberation in groups and the use of scores and grades in evaluation. We study the contributions of these features using a multi-agent computer simulation of groups of physicians. One might expect individual differences in members' grading standards to reduce the capacity of the group to discover the facts on which well-informed decisions depend. Observations of the simulated groups suggest on the contrary that this kind of diversity can in fact be conducive to epistemic performance. Sometimes, it is adopting common standards that may be expected to result in poor decisions.

## Introduction

It has long been known that collecting together inputs from several people can increase the probability of correct decisions and the accuracy of judgments [1,2]. More recently, it has become clear that including different perspectives and ways of thinking within the group can be just as important as individual expertise [3,4]. These insights provide an epistemological rationale for much practice that has evolved in medicine, where knowledge is pieced together in morning reports, case conferences, peer to peer consultations and in other meetings and discussions among staff. Improvements in the quality and efficiency of healthcare may be expected to result from increased reliance on collective intelligence in medicine [5].

Scores and grades play a central part in medical evaluation. For example, Wells and Geneva scores are used in clinical medicine to determine pre-test probabilities for pulmonary embolism, expressed in the qualitative probability grades low, intermediate and high [6,7]; decisions about further diagnostic testing depend on the outcome.<sup>1</sup> Following the Grading of Recommendations Assessment, Development and Evaluation (GRADE) method, systematic review panels categorize bodies of evidence as high,

moderate, low or very low in quality [8]. From Apgar scoring of newborn babies [9] to the Glasgow Coma Scale used to identify organ donors [10], scores and grades express many different aspects of health throughout our lives.

Using grades and other qualitative language brings an important advantage when decisions are made under pressure of time. It's not possible to know, say, exactly what the probability is that a patient who's just come into the emergency room has a pulmonary embolism. That's not necessary, though, in order to judge that the probability is high, or intermediate, or low, and to decide on that basis whether diagnostic procedures such as a CT scan, a chest X-ray or blood tests will be carried out. The reason is just that qualitative probabilities are coarse grained, each covering a range of precise probabilities. There would appear to be drawbacks to grading as well, though. One is that expressions such as 'high', 'intermediate' and 'low' can mean different things to different people. Large differences between doctors and their patients have been documented in the medical literature (see for instance [11]). Similar differences were found among members of a science panel in the Netherlands [12], and among students of business and the social sciences (Fig. 1, [13]).

One solution to problems arising from different interpretations of grades is to establish standard grading procedures for people to use. Take for instance the Geneva and Wells rules for determining pre-test probabilities. They require scoring patients on relevant

<sup>1</sup>A pre-test probability is the subjective probability that a patient has some given condition, before a diagnostic test result is available.

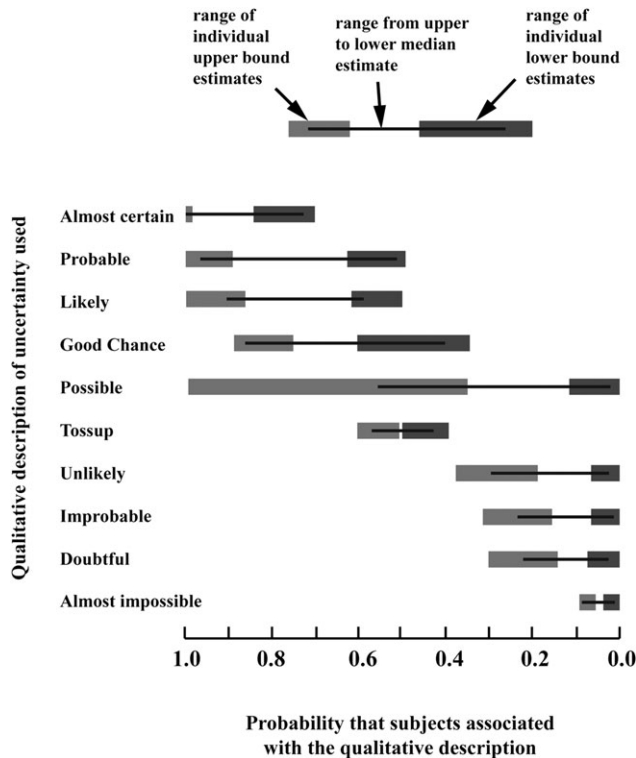


Figure 1 Reproduced from [14], redrawn by [13].

criteria, and then adding up the scores to determine which grade applies ([15], Table 7). Using the GRADE method, bodies of evidence are categorized initially on the basis of study design, and then up-graded or down-graded according to relevant strengths or weaknesses of the study, such as large effect size, or a serious risk of bias ([8], Table 3).

Another approach is to stipulate explicit definitions of grading expressions. In some cases it is possible to specify precise thresholds for applicability, as the Inter-governmental Panel on Climate Change has done for the qualitative probability grades used in its publications [16]. With a multi-faceted notion such as quality of evidence, it is not straightforward to specify thresholds. However, it is possible to characterize grades in ordinary language. For example, the GRADE working group specifies that moderate quality evidence means:

We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different. ([8], Table 2)

Arguably, these methods can be effective in establishing and maintaining common grading standards. Inter-rater reliability studies tend to confirm this [17,18]. The potential benefits are obvious. Following standard protocols makes decisions transparent. Also, having everybody “on the same page” helps people to communicate the knowledge on which good decision-making depends.

Grading standards are conventional. In the end, it’s completely up to us, say, just which precise probabilities will count as low,

which as intermediate and which as high. We argue, though, that they are not merely conventional, in that which standards we settle on can greatly affect our ability to obtain knowledge and make good decisions.<sup>2</sup> Indeed, we argue, some standards are actually harmful, in that a group that has adopted them will be worse at “tracking the truth”, even, than a group whose members could have any understandings at all of what the applicable grades mean. Furthermore, we argue, which standards are better than which, and which standards are harmful, can be hard to tell. This is because it can turn on the nature of the facts that are to be discovered. For example, when these concern the probabilities of several disorders that a patient might have, it can turn on what these probabilities happen in fact to be.

The capacity of a group to discover facts naturally depends on much more than just its conventions about language and the nature of these facts. The level of expertise of individual group members also matters. So does the size of the group. To help us study complex interactions between these and other factors, we have built a computer model. It simulates a group with the task of ranking some given possibilities in order of their probability, on the basis of opinions expressed by individual group members using qualitative probability grades. The group’s epistemic performance in this task is reckoned as the frequency with which an event that the group judges to be most probable really is that; and by setting different parameters we can observe the consequences of the group’s adopting various conventions about the meaning of the grades in which individual inputs are expressed. These observations are the basis for our claims about the epistemic merits and demerits of grading standards.

We proceed as follows. The first section explains our model of grading in groups and certain assumptions of its implementation. The next section compares the performance of simulated groups of graders working with different grading standards. The final section summarizes the conclusions we draw from these results.

## The model

In this section, we focus on a simple and idealized example of collective scoring and grading from clinical medicine. First, we build a model of this example, based on the Arrow-Sen framework in social choice theory [19,20]. Then, we study the model with the aid of computer simulations. An important innovation is that our model includes the language of grades in which individual people express their judgments, along with their interpretations of this language. Taking interpretations of language into account is necessary for a proper understanding of collective decisions more generally, not just in the present medical context [21].

Here is the example. There is a group of three physicians, and a patient who presents with certain symptoms: chest pains and shortness of breath. There are several possible disorders: (H) the patient is hyperventilating, or (HA) is having a heart attack, or (P) has pneumonia, or (PE) has a pulmonary embolism, or... The physicians are to rank the different ones, on the basis of the

<sup>2</sup>Similarly, legal speed limits are not merely conventional. They have consequences for public safety, levels of pollution and the efficiency of our transport system. It’s different with a simple matter of coordination such as which side of the road we drive on.

available information, in order of their pre-test probabilities.<sup>3</sup> The goal is to identify the disorder with the highest pre-test probability.<sup>4</sup>

We now describe a mathematical model of this case. Let  $N$  be the set of physicians, e.g.  $N = \{\text{Brown, Jones, Smith}\}$ , and let  $D = \{d_1, d_2, \dots, d_n\}$  be the set of possible disorders. The physicians will evaluate these diagnoses using a grading language  $L = (G, \succ, I_i)$ . The first component  $G$  is the set of available grades. For example, we might have

$$G = \{high, medium, low\}.$$

The second component  $\succ$  is a linear ordering of the grades. The obvious ordering for these three is:  $high \succ medium \succ low$ .

The interpretation function  $I_i$  tells us, intuitively speaking, how each physician interprets all of the grades. For any given  $n \in N$ ,  $I_n$  assigns to each  $g \in G$  an interval of probabilities between 0% and 100%; this interval is  $n$ 's interpretation of  $g$ . For example,  $I_{\text{Brown}}$  could be the function such that

$$I_{\text{Brown}}(high) = [66\%, 100\%]$$

$$I_{\text{Brown}}(medium) = [33\%, 66\%]$$

$$I_{\text{Brown}}(low) = [0\%, 33\%]$$

Then Brown understands the grade *high* to correspond to a pre-test probability from 66% to 100%, and so on. Of course, different people may have different understandings. For example, Jones might interpret the grades as follows:

$$I_{\text{Jones}}(high) = [90\%, 100\%]$$

$$I_{\text{Jones}}(medium) = [10\%, 90\%]$$

$$I_{\text{Jones}}(low) = [0\%, 10\%]$$

If  $I_m = I_n = I$ , for every  $m, n \in N$ , then  $I$  is called a *consensus interpretation*. Intuitively, with a consensus interpretation, everyone in the group has the same understanding of what the

<sup>3</sup>The pre-test probability  $P(D^+)$  of a disorder is understood here to be the proportion  $D^+ / (D^+ + D^-)$  of patients with this disorder among all those with this particular patient's symptoms.

<sup>4</sup>We assume that identifying the most probable condition is relevant to medical decisions. Sometimes, it certainly is: in determining the Wells score for PE, for instance, physicians are expected to determine whether PE is the most probable diagnosis – or, anyway, whether it is at least as likely as any other [6]. In general, perhaps, it is more important instead to identify those possibilities that require immediate action; for now, though, we assume that what matters is simply probability.

grades mean. Such a consensus might arise naturally, or else it might be the result of everyone's having been told to interpret the grades in some particular way.

In practice, physicians often (but not always) use scoring systems, such as the Wells and Geneva systems, to assign grades to possible diagnoses. We could model these systems explicitly. Instead we take a more general approach that abstracts away from the details of whatever scoring system or other process it is by which physicians arrive at grades. We think of it like this: physicians are presented with some information,  $i$ . It includes the patient's symptoms, medical history and so on. For each disorder  $d$ , they convert this information  $i$  into a corresponding grade  $g$ . Mathematically, we can think of this process as the following mapping:

$$(i, d) \rightarrow g$$

This mapping might be instantiated via a scoring system that the physician has been instructed to use or it might be instantiated by some other process – for example, the physician might make an intuitive judgement on the basis of  $i$  that the pre-test probability of  $d$  is  $g$ .

In order to have a model that we can study systematically, we now add some details. First, we assume that  $i$  and  $d$  together fix some correct pre-test probability  $p \in [0, 1]$ . Second, we introduce the notion of an individual point estimate. The idea is that for any given individual  $n$  there is some noisy process that turns  $p$  into  $n$ 's estimate  $e$  thereof. For simplicity, we assume this process is an unbiased Gaussian with variance  $\sigma^2$  (i.e. a normal distribution or 'bell curve'). Third, we assume that  $e$  is then converted to the appropriate grade  $g$ , according to  $n$ 's interpretation  $I_n$ . For example, if  $e = 5\%$  and we use Jones' above interpretation, then the grade Jones assigns to the disorder in question is *low*. The model, then, can be thought of as the following sequence of mappings:

$$(i, d) \rightarrow p \rightarrow e \rightarrow g$$

Note that we do not assume that  $n$  is *aware* of  $e$  or actually constructs  $e$  in any conscious way. Indeed, we think that in many situations this is rather unlikely (scoring systems typically do not mention point probabilities). Instead, we take the above to be an abstract model of individual grading. As such,  $e$  is a theoretical parameter that we use to generalize across many different ways of moving from  $(i, d)$  to  $g$ , including the use of scoring rules and other protocols that make no reference to point probabilities or estimates.

Based on this noisy process and the individual interpretations of the grades, the individuals assign grades to each of the disorders. Each individual then selects the disorder they give the highest grade. (We must allow for the possibility that several disorders get the top grade. For example, Brown might think that both pneumonia and pulmonary embolism are high probabilities. In such situations, we assume that ties are broken by randomly selecting from the top grade category.) Either the individual's selection is correct, being at the top of the objective ordering, or

else it is not. The individual's expected performance is, by definition, the proportion of cases in which the selection is correct, within a large number of trials.

So far, we have been busy with individual judgements. However, we are interested in the judgements of the group. These are obtained by aggregating the individually assigned grades. Different procedures can be used for this. Here, we follow the simple method of taking the median or middlemost of the individually assigned grades: the grades assigned to  $d$  by the different  $i \in N$  are first listed in order of  $>$  (including repetitions, when several have assigned the same grade). If the number of  $N$  is odd, the collective grade for  $d$  is the one that's in the middle of the list. If it is even, the collective grade for  $d$  is randomly selected from the two grades in the middle of the list.<sup>5</sup>

The group chooses the disorder to which it gives the highest grade. (Again, ties are broken by random selection.) Either the group's selection is correct, in that this disorder really is one of those with the highest pre-test probability, or else it is not. The group's expected performance is, by definition, the proportion of cases in which its selection is correct, within a large number of trials.

In this paper, we focus on individual interpretations of the probability expressions. Would it be good for the group to have a single common interpretation – a semantic consensus, as we will call it? If so, which semantic consensus will be best, in the sense that it is most favourable for determining pre-test probabilities? Even in our simple example, this depends in complicated ways on many factors. In order to answer such questions as these, we take up a tool that is often used to study complex social interactions: multi-agent computer simulation [22]. That requires making further modelling assumptions, to which we now turn.

For one thing, we need an assumption about how the pre-test probabilities of the different possible disorders are distributed. We don't assume the disorders are mutually exclusive: a patient can have several at once. For simplicity and for some degree of realism, we assume that the pre-test probabilities are distributed according to a Pareto distribution,  $\alpha x^\alpha / x^{\alpha+1}$ , where  $\alpha$  is a parameter that we can vary in the simulations.<sup>6</sup> For each simulation run of our model, a distribution of 30 pre-test probabilities is generated from a Pareto sample; the individuals then grade the diagnoses according to the model described above with  $\sigma^2 = 10\%$ , and then their performance is assessed in terms of how often they choose disorders with the highest pre-test probability. Finally, the group grades are determined by aggregating the individual grades and the group performance is then also assessed in terms of how often it chooses the most probable disorders. Performance scores are averaged over 2000 simulation runs.

Regarding the grades and interpretations, we assume that the physicians have 3 grades available to them (this is realistic: many real scoring and grading protocols use from 2 to 4). And in our

simulations, we focus on the three kinds of consensus interpretations and one non-consensus set of interpretations:

- 1 A symmetric consensus interpretation with thresholds of [33, 66, 100]. (That is, individuals with this consensus interpret *low* as [0, 33], *medium* as [33, 66] and *high* as [66, 100].)
- 2 A bottom-heavy consensus interpretation with thresholds of [25, 50, 100].
- 3 A top-heavy consensus interpretation with thresholds of [50, 75, 100].
- 4 A random set of interpretations, where each individual has thresholds  $[a, b, 100]$ , where  $a$  and  $b$  are generated by randomly selecting uniformly from the (0, 100) interval.

Our goal here is not to study every conceivable interpretation, but rather to demonstrate how dramatically the performance associated with a particular consensus can turn on the details of the situation at hand. Similarly, in making our many precise and idealizing assumptions – for example, that the pre-test probabilities are approximately Pareto – we aim to illustrate how the performance associated with a particular consensus can be highly context sensitive. The upshot is that when designing scoring and grade systems, we ought to think carefully about the details of the situations in which they are going to be employed. Computer simulations, we argue, can help us greatly with that.

## Results

We studied 2000 simulation runs of the model described in the previous section for two parameter settings of the Pareto distribution from which the pre-test probabilities are sampled:  $\alpha = 3$  and  $\alpha = 20$ . These are summarized in Figs 2 and 3. Recall that smaller values of  $\alpha$  make more disorders more likely. So, we have effectively studied two kinds of situations: one in which the vast majority of disorders are unlikely ( $\alpha = 20$ ) and another in which 50% and some are even quite likely, with probabilities above 80% ( $\alpha = 3$ ).

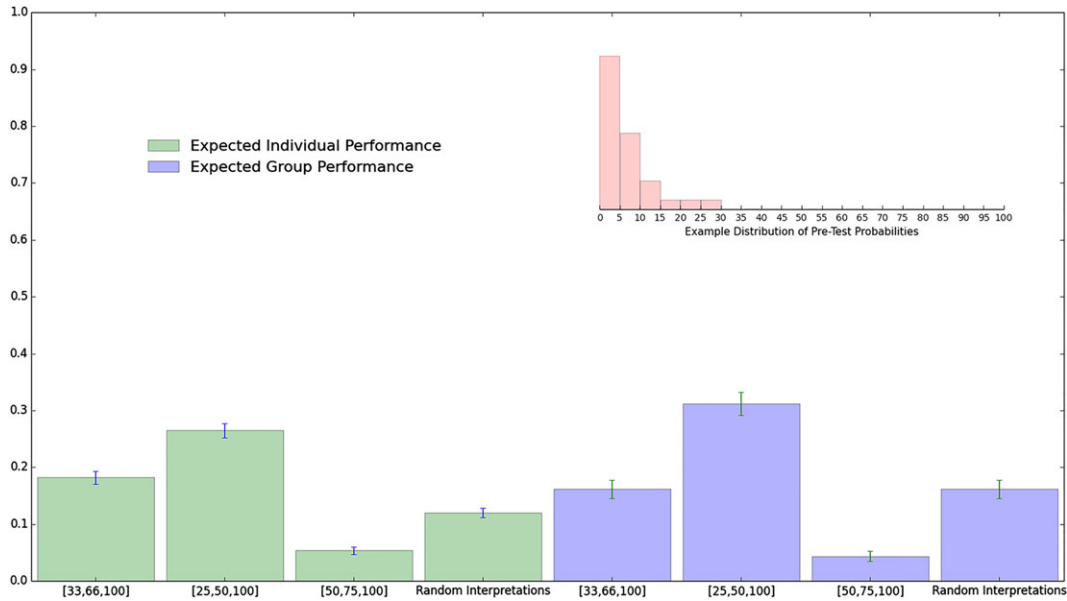
In the first kind of situation (Fig. 2), the bottom-heavy consensus – that is, the '[25,50,100]' consensus – gives the best performance (for both individuals and groups), and the top-heavy consensus – that is, the '[50,75,100]' consensus – gives the worst performance (again, for both individuals and groups).

However, in the second kind of situation (Fig. 3), the exact reverse occurs: the top-heavy consensus gives rise to the best performance (for both individuals and groups), and the bottom-heavy consensus gives rise to the worst performance (again, for both individuals and groups). This reason is this. In the first kind of situation, the bottom-heavy consensus makes more distinctions where most of the pre-test probabilities are – see the example distribution of pre-test probabilities in Fig. 2. And, in the second situation, the top-heavy consensus makes more distinctions where they matter: where the very likely pre-test probabilities are – see the example distribution of pre-test probabilities in Fig. 3.

This reversal demonstrates just how sensitive the performance associated with a particular consensus can be. The only difference between the two kinds of situations is the variation of the parameter  $\alpha$  from 20 to 3. In real life, situations will vary in many more – and much more complicated – ways.

<sup>5</sup>One advantage of making decisions in groups is that people's positions change in light of group discussion. By simply aggregating the individually assigned grades, we leave this important aspect of deliberation to one side. Modeling it is a promising line for future work of the sort begun here.

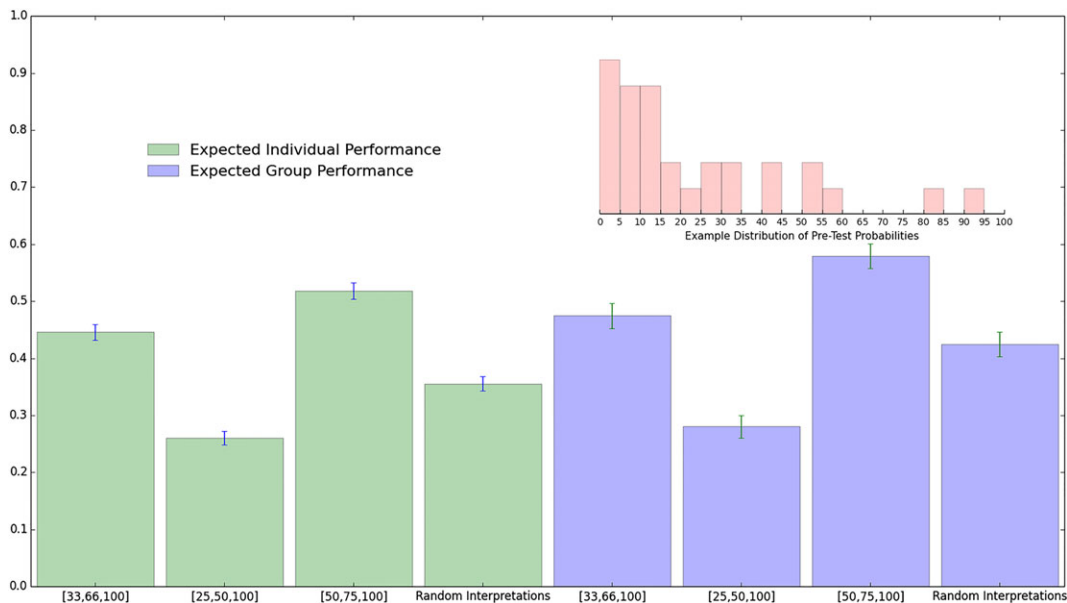
<sup>6</sup>With a Pareto distribution, the different diagnoses can be 'bunched up' at the bottom, with many improbable ones and only a few that are probable. See for example Fig. 2. Smaller values of  $\alpha$  make more diagnoses more likely.



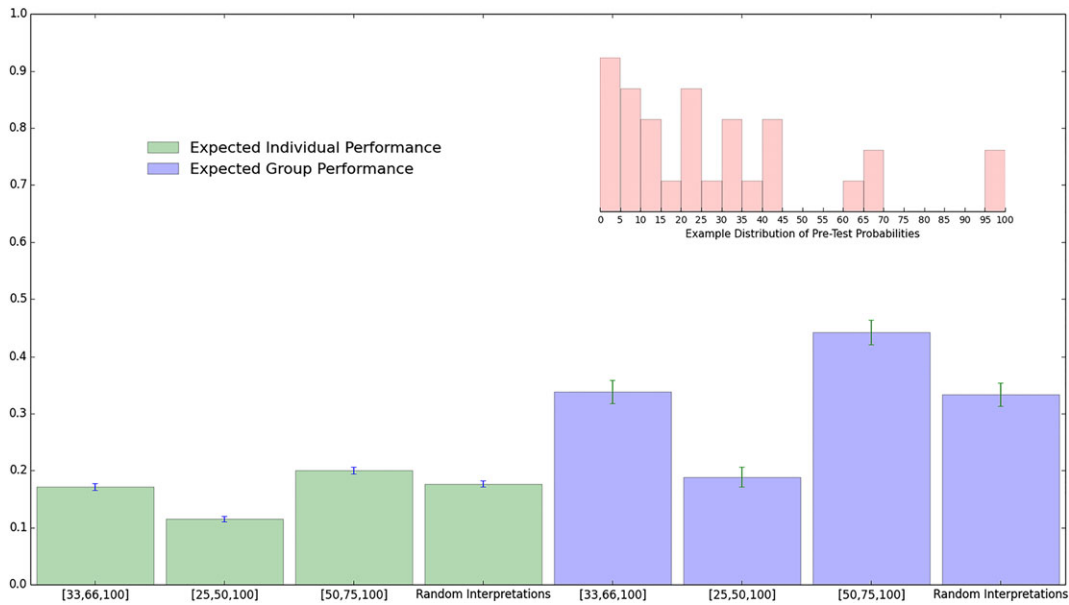
**Figure 2** Performances of different interpretations for individuals (green) and groups (blue) for a Pareto distribution of pre-test probabilities with  $\alpha = 20$ . Error bars are 95% confidence intervals based on 2000 simulation runs.

In both situations, some semantic consensus resulted in worse epistemic performance than random interpretations: in Fig. 2, the top-heavy consensus was worse; in Fig. 3, the bottom-heavy consensus was. This suggests that if time and energy are to be spent on forming a semantic consensus, it is important to think carefully about which it will be; otherwise, it might well be better simply to let everyone interpret the grades just as they please. Surprisingly, perhaps, letting people understand grades any old how can result in better performance than reaching a common understanding.

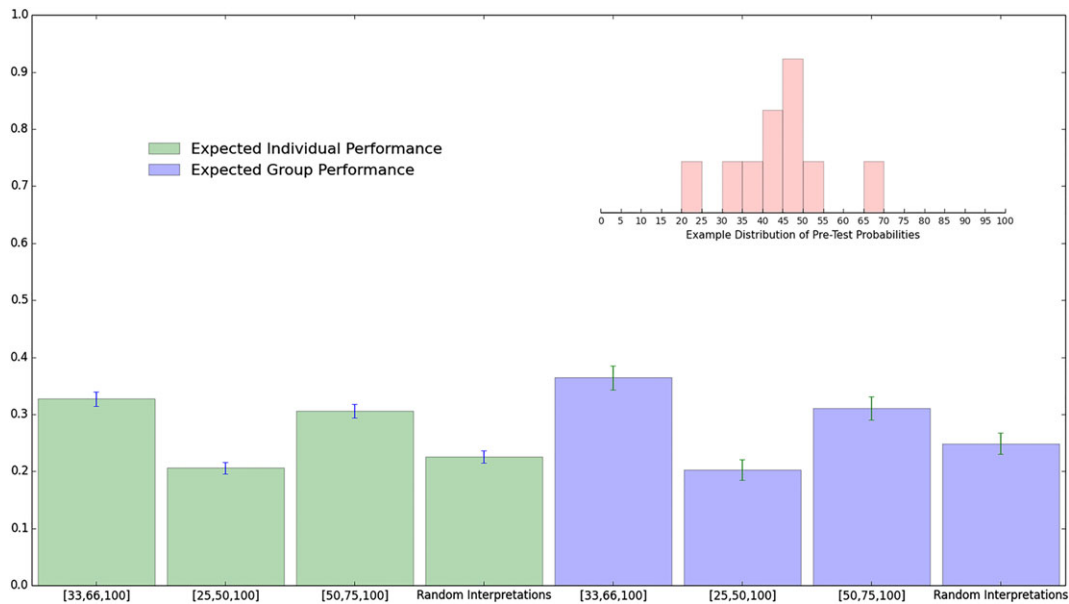
In both situations, we see collective wisdom effects – that is, group performance tended to be better than individual performance – but they are small. This is mostly because the group of physicians was assumed to be small, with just three members, and because the noise associated with the process of generating the hypothetical estimates was fairly low ( $\sigma^2 = 10\%$ ). Clearer differences between individual and group performance can be observed with more noise (e.g.,  $\sigma^2 = 30\%$ ) and a larger group (e.g.,  $|M| = 10$ ). This is demonstrated in Fig. 4.



**Figure 3** Performances of different interpretations for individuals (green) and groups (blue) for a Pareto distribution of pre-test probabilities with  $\alpha = 3$ . Error bars are 95% confidence intervals based on 2000 simulation runs.



**Figure 4** Performances of different interpretations for individuals (green) and groups (blue) for a Pareto distribution of pre-test probabilities with  $\alpha = 3$ , noisier physicians ( $\sigma^2 = 30\%$ ) and more physicians ( $|M| = 10$ ). Error bars are 95% confidence intervals based on 2000 simulation runs.



**Figure 5** Performances of different interpretations for individuals (green) and groups (blue) for a Gaussian distribution with mean 50%, variance 10% and 10 possible disorders. Error bars are 95% confidence intervals based on 2000 simulation runs.

In the cases we have looked at so far, there is always a consensus interpretation that performs better than the symmetric consensus. However, by changing some of the assumptions of our model, we find situations in which the symmetric consensus performs best (of the interpretations we discuss here). For example, changing the underlying distribution of pre-test probabilities to a Gaussian, with

a mean of 50% and a variance of 10%, and reducing the number of possible disorders from 30 to 10 (Fig. 5), we observe that the symmetric consensus performs better than the other consensus interpretations and random interpretations. Again, we see how the relative performance of a consensus interpretation is sensitive to the details of the situation at hand.

## Summary and discussion

People often find it easier to grade, and to communicate with others in qualitative terms, rather than to use more rigorous language. In order to avoid interpersonal differences in the interpretation of grades, steps can be taken to standardize usage. These include the use of scoring and grading protocols, and providing explicit definitions of the relevant expressions.

We showed using a computer simulation that, from an epistemological viewpoint, some standards are better than others. Importantly, some standards can result in very poor performance, in that a group having adapted them may be expected to be less capable of discovering relevant facts, even, than a group of people whose members could interpret the grades in any way at all. (This can be seen in Figs 2 & 3, where the performances of random interpretations were better than top-heavy and bottom-heavy interpretations, respectively.) Crucially, which standards lead to better performance, and whether adopting any given standards would help the group or hurt it, is highly contextual. In the case we considered here, it depends, among other things, on what the probabilities of the given disorders happen to be.

Our model of individual and collective grading is very simple. Surely some of its assumptions and idealizations will be found unrealistic. The task we analyzed, choosing the most-probable disorder, is also simple. A more important task, perhaps, is choosing the disorder with the highest expected danger, because although such disorders might be relatively unlikely, they can require immediate action. Because some disorders with substantial danger are more likely than others, this should have a significant impact on the performances associated with different grading standards. With multi-dimensional tasks, there is also further scope for interpersonal differences and for standardization, because in addition to having different thresholds, people can attach different weights and priorities to the several dimensions. There is much work to be done before our simulation studies cover any large part of real medical practice.

Be this as it may, the results already point to the contribution that computational studies can make. Scores, grades and other qualitative language are used throughout medicine, for estimating probabilities, evaluating evidence and more. Our simulations show not only that different ways of standardizing such language can have very different consequences for performance, but also that many other factors are relevant, all interacting in complicated and sometimes unintuitive ways. Introducing computational studies and statistical analysis into the development of scoring and grading systems may be expected to help researchers to think through the consequences of basic design decisions such as how many grades to use, how many people are needed for expert panels to be effective and what the required levels of expertise are – before their decisions become well-established in medical practice.

## References

1. Condorcet, J. A. N. d. C. (1785) *Essai sur l'application de l'analyse a la probabilité des décisions rendues a la pluralité des voix* [microform] / par M. le Marquis de Condorcet. Paris: Imprimerie royale Paris.
2. Galton, F. (1907) *Vox Populi*. *Nature*, 75, 450–451.
3. Page, S. (2008) *The Difference*. Princeton, NJ: Princeton University Press.
4. Surowiecki, J. (2004) *The Wisdom of Crowds*. New York, NY: Doubleday.
5. Wolf, M., Krause, J., Carney, P. A., Bogart, A. & Kurvers, R. H. (2015) Collective intelligence meets medical decision-making: the collective outperforms the best radiologist. *PLoS One*, 10 (8), e0134269.
6. Wells, P. S., Ginsberg, J. S., Anderson, D. R., et al. (1998) Use of a clinical model for safe management of patients with suspected pulmonary embolism. *Annals of Internal Medicine*, 129 (12), 997–1005.
7. Wicki, J., Perneger, T., Junod, A., Bounameaux, H. & Perrier, A. (2001) Assessing clinical probability of pulmonary embolism in the emergency ward: a simple score. *Archives of Internal Medicine*, 161 (1), 92–97.
8. Balshem, H., Helfand, M., Schunemann, H. J., et al. (2011) GRADE guidelines: 3. Rating the quality of evidence. *Journal of Clinical Epidemiology*, 64 (4), 401–406.
9. Finster, M. & Wood, M. (2005) The Apgar score has survived the test of time. *Anesthesiology*, 102 (4), 855–857.
10. Teasdale, G. & Jennett, B. (1974) Assessment of coma and impaired consciousness: a practical scale. *The Lancet*, 304 (7872), 81–84. Originally published as Volume 2, Issue 7872.
11. Ohnishi, M., Fukui, T., Matsui, K., et al. (2002) Interpretation of and preference for probability expressions among Japanese patients and physicians. *Family Practice*, 19 (1), 7–11.
12. Wardekker, J. A., van der Sluijs, J. P., Janssen, P. H. M., Kloprogge, P. & Petersen, A. C. (2008) Uncertainty communication in environmental assessments: views from the Dutch science-policy interface. *Environmental Science & Policy*, 11, 627–641.
13. Morgan, M. G. (2014) Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National Academy of Sciences of the United States of America*, 111 (20), 7176–7184.
14. Wallsten, T., Budescu, D., Rapoport, A., Zwick, R., Forsyth, B. (1986) Measuring the Vague Meanings of Probability Terms. *Journal of Experimental Psychology: General*, 115 (4), 348–365.
15. Torbicki, A., Perrier, A., Konstantinides, S., et al. (2008) Guidelines on the diagnosis and management of acute pulmonary embolism. *European Heart Journal*, 29 (18), 2276–2315.
16. Mastrandrea, M., Mach, K., Plattner, G. K., Edenhofer, O., Stocker, T., Field, C., Ebi, K. & Matschoss, P. (2011) The IPCC AR5 guidance note on consistent treatment of uncertainties: a common approach across the working groups. *Climatic Change*, 108 (4), 675–691.
17. Iles, S., Hodges, A., Darley, J., Frampton, C., Epton, M., Beckert, L. & Town, G. (2003) Clinical experience and pre-test probability scores in the diagnosis of pulmonary embolism. *QJM*, 96 (3), 211–215.
18. Mustafa, R. A., Santesso, N., Brozek, J., et al. (2013) The GRADE approach is reproducible in assessing the quality of evidence of quantitative evidence syntheses. *Journal of Clinical Epidemiology*, 66 (7), 736–742.e5.
19. Arrow, K. J. (1951) *Social Choice and Individual Values*, Second edn. New York: Wiley, 1963.
20. Sen, A. K. (1970) *Collective Choice and Social Welfare*. San Francisco: Holden-Day.
21. Morreau, M. & Weymark, J. A. (2016) Measurement scales and welfarist social choice. *Journal of Mathematical Psychology*. In press. doi:10.1016/j.jmp.2016.04.001.
22. Epstein, J. M. (2007) *Generative Social Science: Studies in Agent-Based Computational Modeling*. Princeton, NJ: Princeton University Press.