

**An appropriate verbal probability lexicon for communicating surgical risks is *unlikely*
to exist**

Adam J. L. Harris¹, Tracy Tran¹, Sarah C. Jenkins¹, Adelia Su¹, Lexi He¹, Yifei Zhu¹ and
Simon Gane²

¹Dept. of Experimental Psychology, University College London

²Royal National Ear, Nose and Throat and Eastman Dental Hospital

Author Note

Adam J. L. Harris, Tracy Tran, Sarah Jenkins, Adelia Su, Lexi He and Yifei Zhu,
Dept. of Experimental Psychology, 26 Bedford Way, London, UK, WC1H 0AP. Simon
Gane, Royal National Ear, Nose and Throat and Eastman Dental Hospital, 47-49 Huntley St,
London, WC1E 6DG.

Sarah Jenkins is now at Dept. of Psychology, Royal Holloway, University of London.

We thank the Pan-Thames ENT Specialist Registrars who generously gave us some of
their time.

The materials and data from this project can be accessed here:
https://osf.io/yut2g/?view_only=270fec2f50484250a5a0d56bafb3496d

Correspondence concerning this article should be addressed to Adam J. L. Harris
(email: adam.harris@ucl.ac.uk).

Abstract

Effective risk communication about medical procedures is critical to ethical shared decision-making. Here, we explore the potential for development of an evidence-based lexicon for verbal communication of surgical risk. We found that Ear, Nose and Throat (ENT) surgeons expressed a preference for communicating such risks using verbal probability expressions (VPEs; e.g., 'high risk'). However, there was considerable heterogeneity in the expressions they reported using (Study 1). Study 2 compared ENT surgeons' and laypeople's (i.e., potential patients) interpretations of the ten most frequent VPEs listed in Study 1. Whilst both groups displayed considerable variability in interpretations, lay participants demonstrated more, as well as providing systematically higher interpretations than those of surgeons. Study 3 found that lay participants were typically unable to provide unique VPEs to differentiate between the range of (low) probabilities required. Taken together, these results add to arguments that reliance on VPEs for surgical risk communication is ill-advised. Not only are there systematic interpretational differences between surgeons and potential patients, but the coarse granularity of VPEs raises severe challenges for developing an appropriate evidence-based lexicon for surgical risk communication. We caution against the use of VPEs in any risk context characterised by low, but very different, probabilities.

Keywords: ENT; informed consent; risk communication; verbal probability expressions; surgical risk

Public significance statement

ENT (Ear, Nose and Throat) surgeons report most often communicating surgical risks with verbal probability expressions (VPEs, e.g., ‘unlikely’; ‘likely’), but their interpretations of oft-used terms differ significantly from those of laypeople (Studies 1 & 2). In Study 3, laypeople were unable to suggest VPEs to differentiate ten numerical probabilities inferred (from surgeon’s data in Studies 1 and 2) to be relevant for surgical risk communication. We propose that communicating low probability risks, as required in the surgical domain, is best done with numbers to enable appropriate differentiation between risk levels.

An appropriate verbal probability lexicon for communicating surgical risks is *unlikely* to exist

Informed consent is a fundamental principle in medicine. Patients should be informed of the risks and benefits associated with a surgical procedure, in order to subsequently make an informed decision about whether or not to undergo surgery (General Medical Council, 2008). Similarly, in pharmacy, patients should decide whether or not to take a particular medication in full knowledge of its likely benefits and potential side effects. Conveying the probability of adverse outcomes from surgery or medication would appear best achieved using numbers (as recommended in Trevena et al., 2013). Numbers represent a clear quantitative metric for expressing probability, and reduce the potential for an ‘illusion of communication’, which can arise from the use of verbal probability communications (Budescu & Wallsten, 1995, p. 299). Such an ‘illusion’ arises where the understanding of communication recipients does not match that intended by the communicators using the terms.

There is, however, evidence that communicators generally (Erev & Cohen, 1990), and medical professionals in particular (e.g., Brun & Teigen, 1988), prefer to communicate probabilities verbally – with verbal probability expressions (VPEs; e.g., stating “X is unlikely” rather than “There is a 20% chance of X”). In addition to a preference for verbal communications of probability, some domains prescribe that probabilistic information *should* be communicated in verbal form (e.g., Intergovernmental Panel on Climate Change [IPCC; Mastrandrea et al., 2010]; Intelligence [College of Policing, n.d.; ODNI, 2007; NATO, 2016, as cited in Dhimi & Mandel, 2021]; Pharmacy [MHRA, 2005]). These domains have additionally prescribed specific lexicons for probability communication. Perhaps the most well-known of these comes from the IPCC (Mastrandrea et al., 2010), and is presented in Table 1. Similar lexicons have also been developed by intelligence organisations (e.g., College of

Policing, n.d.; ODNI, 2007; NATO, 2016, as cited in Dhimi & Mandel, 2021), as well as by the European Food Safety Authority (Hart et al., 2019), despite the latter's recommendation to always express uncertainty numerically. In the medical domain, the EU suggested grouping medicine side effect frequencies into five groups, each represented by a single VPE¹ (MHRA, 2005; Table 2). A comparison of Tables 1 and 2 demonstrates context differences in the use and understanding of verbal expressions of risk. A risk of greater than 10% can be described as 'very common' (Table 2) or 'very unlikely' (Table 1). Navigating a disparate variety of VPE interpretations is likely to be a complex task for individuals. Such a view is supported by evidence that laypeople typically overestimate the frequency expressions endorsed by the EU, both in English (Berry et al., 2002, 2003; Knapp et al., 2001; Webster et al., 2017), and in German (Ziegler et al., 2013).

Table 1.

Likelihood scale of the Intergovernmental Panel on Climate Change (IPCC).

Verbal expression	Likelihood of the Outcome
Virtually certain	99-100%
Very likely	90-100%
Likely	66-100%
About as likely as not	33 to 66%
Unlikely	0-33%
Very unlikely	0-10%
Exceptionally unlikely	0-1%

¹ Technically, these are verbal expressions of *frequency*, rather than verbal *probability* expressions (VPEs). Throughout the current manuscript, we use VPE as a common term to represent any verbal expression of either frequency or probability. All uses of VPE in the present manuscript refer to a risk communication.

Table 2*EU likelihood scale for the expression of risk in patient information leaflets*

Verbal expression	Frequency
Very common	>10%
Common	>1% and <10%
Uncommon	0.1% to 1%
Rare	0.01% to 0.1%
Very rare	Up to 0.01%

Note: MHRA (2015, p. 161) recommend such verbal expressions are “only used if accompanied by the equivalent statistical information. For example, ‘Very rarely (fewer than 1 in 10,000 patients treated)...’”

As we have already alluded to, organisational prescriptions to use VPEs are not without their critics. Numerous documentations of interpersonal variance in interpretations (e.g., Beyth-Marom, 1982; Brun & Teigen, 1988; Budescu & Wallsten, 1995; Wallsten et al., 1993; Theil, 2002) suggest that intended probabilities will often not match those understood by the communication recipient. There is evidence that patients’ risk perceptions are higher in response to verbal communications of risk than numerical communications (Berry et al., 2003), and the severity of the outcome being described can influence how VPEs are understood (Bonnefon & Villejoubert, 2006; Harris & Corner, 2011; Holtgraves & Perdeu, 2016; Juanchich et al., 2012; Weber & Hilton, 1990). These are just three potential obstacles to efficient communication associated with the use of VPEs (for reviews see Collins & Hahn, 2018; Mandel et al., 2021). Within a medical context, such phenomena can be highly consequential for the notion of informed consent. The patient’s understanding of harm probability might be very different from the one communicated, thus undermining the *informed* element of informed consent. Such misunderstandings might lead to patients to

accept treatments they would otherwise refuse, or refuse treatments they would otherwise accept (e.g., because of an exaggerated perception of surgical risk).

Given the oft-documented limitations of verbal communications, in the present paper we aimed to: 1) ascertain the prevalence of VPE use amongst Ear, Nose and Throat surgeons; 2) measure the interpersonal (in)consistency in interpretations amongst both surgeons and laypeople (potential patients); 3) develop an evidence-based lexicon to increase the consistency of interpretations (c.f. Ho et al., 2015; Merz et al., 1991). Our focus is on Ear, Nose and Throat (ENT) surgery, although we assume that our results are generalizable across medical domains. In the following, we provide brief reviews of studies investigating: 1) medical professionals' risk communication format preferences; 2) interpretations of VPEs within a medical context.

Previous investigations of medical professionals' risk communication

Format preferences

Previous research has investigated physicians' use of different risk communication formats, as well as patients' understanding of those communications. A number of studies have employed self-report methods. Brun and Teigen (1988), and Juanchich and Sirota (2020), found General Practitioners (GPs) reported a preference for expressing risk verbally rather than numerically (76% and 67% of participants respectively). Anderson et al. (2011) asked obstetrician-gynaecologists to state how they typically communicated risks of Down syndrome following screening test results. 45% reported using words, as opposed to 33% using numbers (the remainder reporting 'other' or it 'varies by test'), although the degree of numerical preference could have been exaggerated due to providing three verbal options, and only a single numerical one. Contrastingly, Ohnishi et al. (2002) found most Japanese GPs (58%) preferred numbers. In qualitative interviews, genetic counsellors reported generally preferring to avoid VPEs, perceiving them as too directive (Henneman et al., 2008). GPs in Petrova et al.'s (2018)

study generally reported that they would communicate more risks associated with cancer screening with numbers than with words (means of 2.1 vs. 1.2 risks [the average number of risks they would communicate with visual aids was 1.1]).

Kunneman et al. (2015), Michie et al. (2005) and Neuner-Jehle et al. (2011) taped consultations between consultants and patients (GPs discussing cardiovascular risk, the benefits and harms of radiotherapy as a treatment for rectal cancer, and genetic consultations respectively). All three studies demonstrated a greater use of verbal over numerical communications of risk. Neuner-Jehle et al. reported that 73% of 70 consultations exclusively communicated risk using VPEs. Kunneman et al. observed that the modal communication format in consultations was to use both verbal and numerical formats (41%), but after that verbal communications were more frequent (33%) than numerical ones (26%). Michie et al. reported that 53% of 492 risk communications were verbal (47% were numerical). Additionally, Pieterse et al. (2006) recorded 51 breast cancer genetic counselling visits. 36% of 419 recorded communications of risk used both VPEs and numbers, with 34% using VPEs alone and 30% using numbers alone.

This brief review reveals some cross-study differences (which there is not yet sufficient evidence to attribute to different medical specialties) in the degree to which medical professionals prefer to communicate risk with VPEs. Nonetheless, across all studies a significant proportion of physicians did prefer to use VPEs. Most of the reviewed studies were, however, published before the first International Patient Decision Aid Standards' (IPDAS) guidance advised that numerical communications should be preferred (Trevena et al., 2013).² In the intervening six years (our first experiment was carried out in late 2019), have preferences changed? Petrova et al.'s (2018) results suggest 'possibly', although GPs in that study were

² Whilst the updated IPDAS guidance (Bonner et al., 2021) echoed this guidance, it also highlighted the importance of future research investigating what risk information might be better communicated without numbers.

themselves presented with explicit numerical risk information and asked how they would present it, potentially exaggerating the reported use of numbers. Juanchich and Sirota's (2020) research, contrastingly, suggests preferences remain for verbal communications, and we subsequently also predict a preference for verbal communications of surgical risk amongst our recruited surgeons. If such a preference is observed, it is important to understand how these expressions are used and understood.

Interpretations of verbal probability expressions

Consistent with research in the general population, medical professionals have been shown to vary in their interpretations of VPEs (e.g., Bryant & Norman, 1980; Merz et al., 1991; Nakao & Axelrod, 1983; Stheeman et al., 1993; Timmermans, 1994). Additionally, medical professionals' (doctors attending a rheumatology conference in the UK) understanding of VPEs describing medicine side effect frequency have been shown to be systematically higher than those intended by both the EU, and the former UK Chief Medical Officer (Calman, 1996; Berry et al., 2004). These overestimates have, however, been shown to be less extreme than those observed in lay participants (Berry et al., 2004; see also Wiles et al., 2020). More generally, lay participants have consistently been shown to differ from medical professionals in their interpretations of VPEs in a medical context (Brun & Teigen, 1988). Lay participants typically demonstrate a more regressive pattern of interpretations (overestimating rare labels, and underestimating common labels) in comparison to medical professionals (Ohnishi et al., 2002; Shaw & Dear, 1990). In studies that have investigated solely laypeople's interpretations of VPEs, the consistent finding is one of considerable interpersonal variability (e.g., Berry et al., 2002; Kunneman et al., 2020; Sutherland et al., 1991).

In the above, we have reviewed studies as they addressed medical professionals *generally*, or laypeople. We are aware of no previous research that has investigated the

interpretation of VPEs in the context of ENT surgery specifically. Wiles et al. (2020) did, however, request medical professionals (anaesthetists and surgeons) and patients awaiting surgery to provide a numerical (percentage) interpretation of seven VPEs (six from Calman, 1996: ‘negligible’, ‘minimal’, ‘low’, ‘moderate’, ‘high’, ‘very high’; as well as ‘standard risk’) specifically referring to the likelihood of a post-operative complication. Consistent with related research, patients displayed greater variance in their interpretations of VPEs, and provided higher numerical translations than did medical professionals (although there was also notable variance in the professionals’ interpretations, especially for ‘high’ [IQR: 10-40%] and ‘very high’ risks [IQR: 20-50%]).

Overview

The heterogeneity (and potential for systematic differences between surgeons and laypeople) in interpretations of VPEs suggests a clear potential benefit of a standardised, evidence-based lexicon (see also Ho et al., 2015; Merz et al., 1991). The current study originally aimed to contribute to such an objective across four studies: 1) VPEs used by ENT surgeons are elicited from them directly; 2) Interpretations of common VPEs elicited from ENT surgeons are compared across surgeons and laypeople, 3) VPEs are elicited from laypeople to represent the numbers required according to the surgeons’ interpretations, and 4) Communication effectiveness is compared across verbal communications that adhere to the expectations of laypeople (from Study 3) versus those VPEs currently utilised (from Study 1).

To foreshadow our results: in the event, we did not run the planned Study 4 (which would have been based on methods pioneered in Budescu & Wallsten, 1990 [see also Erev & Cohen, 1990; Karelitz & Budescu, 2004]). This was because the results of Study 3 suggested that a key desiderata of the evidence-based lexicon that we set out to develop could not be met. Namely, lay participants could not provide an intuitive verbal scale of sufficient granularity to

differentiate the small probabilities relevant in this context. Thus, it was, quite simply, not possible to create verbal communications that matched participants' intuitive use of these terms (in contrast to, e.g., Budescu & Wallsten, 1990, who investigated the communication of probabilities which were all greater than or equal to 5%; see also Erev & Cohen, 1990; Karelitz & Budescu, 2004).

Study 1

Study 1 had three objectives. First, it aimed to determine surgeons' preferences for communicating surgical risks with words versus numbers. Second, it was important to obtain a selection of VPEs that are actually used by ENT surgeons, so as to maximise the relevance of Study 2, where their interpretations would be compared across surgeons and patients (as in Kunneman et al., 2015, 2020). Third, Study 1 sought to determine the heterogeneity in surgeons' choices of VPEs.

Method

Participants

An opportunistic sample of forty-nine ENT Specialist Registrars (SpRs) (23 female), aged 28 – 45 years (median = 33; two participants did not provide their age), volunteered to participate in this survey study. The study was run at a training session during the Pan-Thames (London, Kent, Surrey and Sussex) ENT Training Days (UK) on November 11th, 2019. SpRs are fully qualified medical doctors who continue to undergo higher surgical training in the UK National Health Service (NHS), following four years of post-graduate 'basic' surgical training (two foundation years and two years of core surgical training; see Royal College of Surgeons of England, 2020). The next step in their career path is surgical consultant (for explanation of

medical titles in the UK, see British Medical Association, 2020). The years of surgical training reported ranged from 3³ to 8 ($M = 5.35$, $SD = 1.71$). Forty-four ENT surgeons were native English speakers (non-native speakers were still expert speakers as non-UK medical graduates are required to pass the IELTS [<https://www.ielts.org>] English exam before working in the UK). No incentives were provided for participation. Ethical approval for all studies was from the Departmental Ethics Chair for Speech, Hearing and Phonetic Sciences, University College London (ShaPS_2015_AH_017).

Design and Materials

We were primarily interested in the probability words surgeons reported using, as well as the proportion of instances in which they self-reported using words to communicate surgical risks.

Participants received a six-page questionnaire. The first page requested demographic details. The second page introduced participants to the upcoming tasks, highlighting that risks associated with surgery can be communicated in a variety of ways, but two common ways are using numbers or words. Participants were then provided with examples of numerical communication formats (percentages, frequencies (1 in x), classical probabilities (0-1), odds ratios) and verbal formats, “any **verbal probability or frequency terms** (e.g., *rare*, *unlikely*, *common*)” [emphasis in original]. On p. 3, participants indicated which format (numbers, words, other) they most often used when conveying risk probability to an ENT patient, as well as indicating the percentage of each type of communication that they used. Page 4 asked participants to indicate their preferred communication format in three case studies. These questions were part of a separate study⁴ (Su, 2020) and will not be discussed further here. On

³ SpRs may have less than four years postgraduate *surgical* training if they included some non-surgical disciplines during their foundation training ‘rotations’ (Royal College of Surgeons of England, 2020).

⁴ Participants indicated their preference for communicating risk using words or numbers for side effects with consequences that varied in severity.

p. 5, participants listed “all the verbal probability terms that [they] use when communicating ENT surgery risk with patients.” The final page asked participants to refer back to the previous page and provide a numerical interpretation for each VPE they had written. These data (available at: https://osf.io/yut2g/?view_only=270fec2f50484250a5a0d56bafb3496d) have not been analysed, given the variability in VPEs generated and Study 2’s focussed test of surgeons’ numerical interpretations.

Procedure

Information sheets and consent forms were distributed to participants in a lecture theatre. These were collected, and questionnaires were distributed to all who consented to participate. It took approximately 15 minutes for all questionnaires to be completed, and participants were told that they would be debriefed as to the purpose of the study after the second session (one month later – Study 2).

Results

Risk communication format

The majority of participants (57%) reported communicating risk probability most frequently with words. Only 35% reported most frequently using numbers, with the remaining four participants indicating that they used both. Overall, there was evidence that these surgeons generally preferred words to numbers to communicate surgical risk probability, $\chi^2(2) = 17.7$, $p < .001$.

What VPEs do surgeons report using?

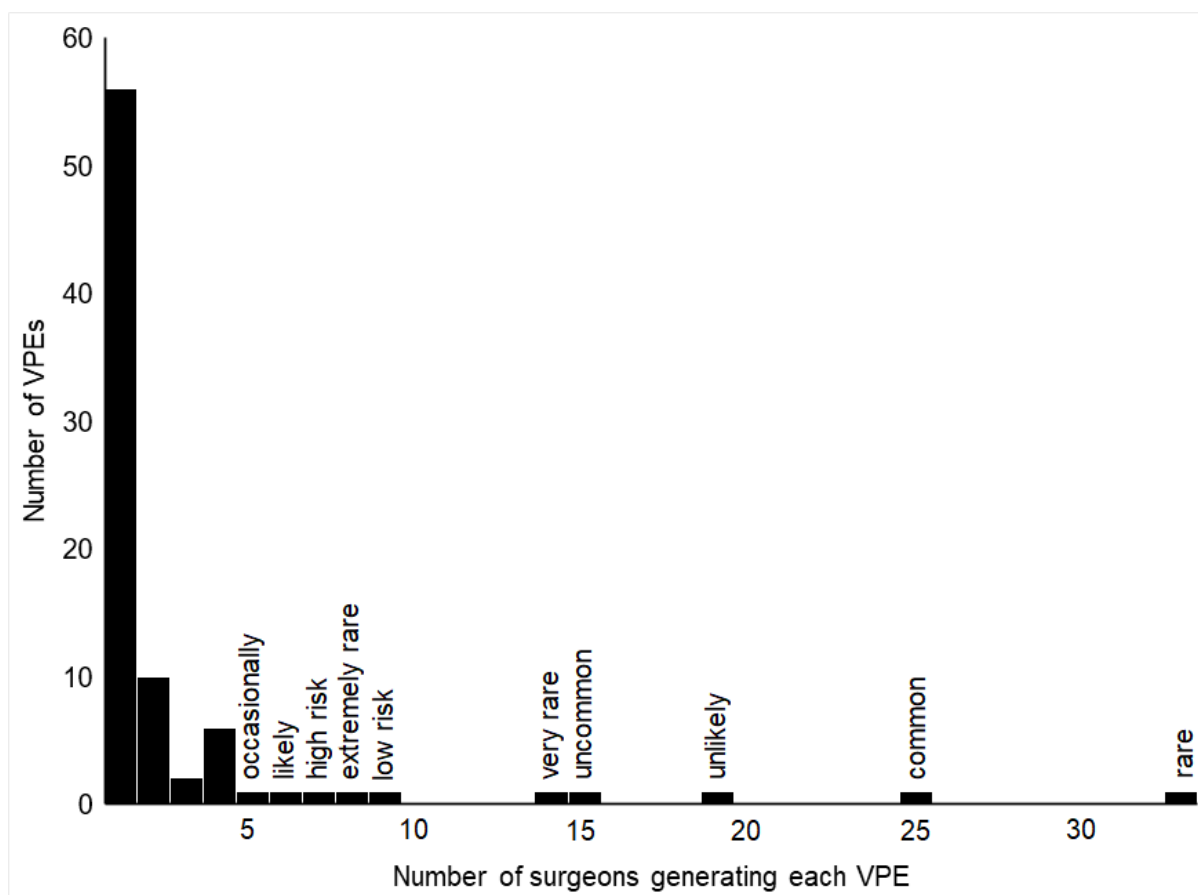
TT and AS coded the data independently and resolved any disagreements in discussion.⁵ In total, 84 different VPEs were generated (see Appendix A⁶). 56 of these VPEs were only generated by a single surgeon (Figure 1). Thus, there is considerable heterogeneity in surgeons' specific VPE use. The most popular VPE was 'rare' (generated by 33 surgeons [67%]). 10 VPEs were generated by five or more surgeons (see Figure 1), and these were used as stimuli in Study 2.

⁵ We do not have a formal record of the agreement statistics, but disagreements were very rare (we recognise the irony of this statement) and were associated with handwriting.

⁶ As can be seen from Appendix A, coding erred towards coding more VPEs as unique – e.g., 'rare' (N = 33) and 'rarely' (N = 4) were coded as distinct expressions.

Figure 1

Number of surgeons generating each VPE in Study 1.



Note. The most frequently generated VPEs were: ‘rare’ (33), ‘common’ (25), ‘unlikely’ (19), ‘uncommon’ (15), ‘very rare’ (14), ‘low risk’ (9), ‘extremely rare’ (8), ‘high risk’ (7), ‘likely’ (6), ‘occasionally’ (5).

Study 2

Study 1 revealed that, whilst most surgeons reported most often using words for risk communication, there was little agreement on the actual VPEs used. For those VPEs most frequently endorsed by surgeons, however, do potential patients interpret them as the surgeons intend? To answer this question, Study 2 tested the consistency between surgeons’ and patients’ understanding of the ten most commonly generated VPEs from Study 1.

Method

Participants

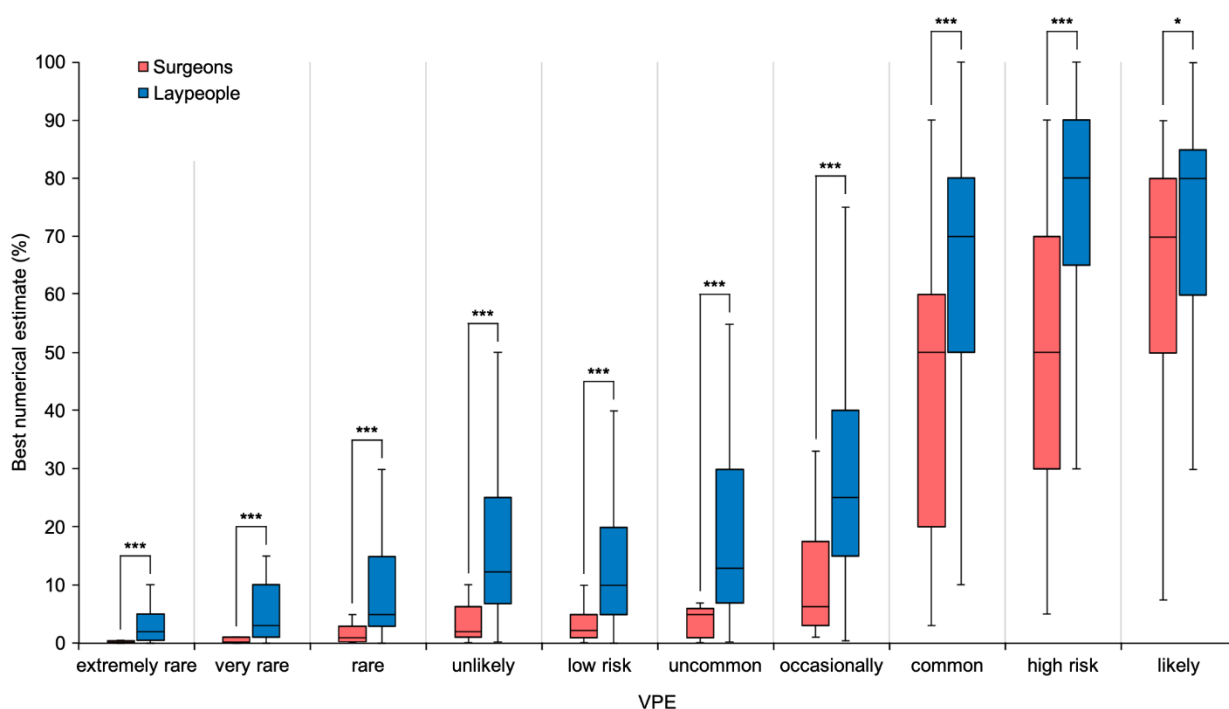
An opportunistic sample of forty-four ENT SpRs (19 female) was recruited. ENT surgeons, aged 28 – 39 years (median = 32), volunteered to participate at a training session (December 17th, 2019) during the Pan-Thames ENT Training Days (of the same cohort, thus there was substantial overlap in the surgeons participating in Studies 1 & 2). The years of surgical training reported ranged from 3 to 8 ($M = 4.90$, $SD = 1.53$; one participant did not answer this question, and one reported not being an SpR. Due to the people present in the room, we assumed this 39 year old was a consultant surgeon, and therefore retained them in analyses). No incentives were provided for surgeons' participation. In an attempt to ensure a similar number of complete datasets across surgeons and laypeople, 60 laypeople (37 female; sample size determined before data collection commenced) were recruited. Lay participants, aged 18 – 76 years (median = 37), were recruited online through Prolific.co and received £0.80 (average payment rate £6.51 per hour). Eighty percent (35) of the ENT surgeons and 88% (53) of the lay participants were native English speakers.

Design

A 10 (VPE) \times 2 (participant group) mixed design was employed with VPE manipulated within participants. Participants provided a lower bound, upper bound and best numerical estimate for each VPE (see Figure 2). The VPEs were presented in one of four different orders. Participants could provide their estimates in any numerical format they preferred.

Figure 2

Best estimates for numerical interpretations of the ten VPEs by both surgeons and laypeople.



Note. The solid line within the box represents the median, the limits of the box represent the interquartile range, and the whiskers represent the range.

* $p < .05$; ** $p < .01$; *** $p < .001$ [by Mann-Whitney U tests].

Materials and Procedure

Surgeons. Surgeons were first given an information sheet and provided informed consent. At the start of the questionnaire, surgeons read: “In this study, we are interested in how you communicate the risks associated with ENT surgery to a patient. This questionnaire concerns the format of communication that you use to convey risks of surgical side effects to patients.”

The critical questions for this study were on p. 5 of a seven page questionnaire (full materials at https://osf.io/yut2g/?view_only=270fec2f50484250a5a0d56bafb3496d). After providing demographic details, participants answered a series of questions pertaining to a

separate study (Su, 2020). On the critical p. 5, surgeons were informed that we were interested in their “numerical interpretations of verbal probability terms,” and that the table below includes “terms that surgeons may use when communicating the risks associated with ENT surgery to a patient.” Surgeons were further informed that because “terms often convey a range of numerical probabilities⁷”, we wanted them to provide a “**lower bound** and **upper bound** for such a range, as well as a **best estimate** for each term” [emphasis in original]. Numerical probabilities can be expressed in a range of formats, with different people having different preferences, which can vary by context and probability level (see e.g., Bonner et al., 2021). Participants were therefore instructed they could use whatever format they believed would provide the best numerical interpretations, and that decimal places were allowed. They were subsequently provided with four example formats: percentages, frequencies (“___ in ___”), classical probabilities (0 to 1) and odds ratios. At the bottom of the page was a table containing the ten VPEs (see Figure 2) in the left hand column, with boxes for lower bound, upper bound and best estimates (in that order) in the three adjacent columns. The final question on the questionnaire asked participants whether they saw 33.3% or 50% as equivalent to a ratio of 1:2. This was because it was felt that some participants might misunderstand odds ratios, which was important to know for analysis purposes. Finally, participants were provided with a written debrief and given the opportunity to ask questions.

Lay participants. Whilst we intended to minimise differences between the materials provided to surgeons and lay participants, some additional instructions were provided to lay participants to ensure understanding of the task (especially necessary given this part of the

⁷ We included this text to provide further context for the request to provide lower and upper bounds. An anonymous reviewer highlighted that this could have added variance to the data, as participants might have interpreted this in a number of different ways, including: ‘the range of ways other people interpret these expressions’ or ‘my subjective range of probabilities.’ We acknowledge this limitation, but also that note that it is unlikely that this would affect comparisons between surgeons and lay people, given that the wording was the same for both groups of participants.

study was conducted online). In addition, we required lay participants to put themselves in the position of patients (surgeons were in the position of surgeons).

Upon direction to the study website, lay participants provided informed consent, where the ENT context was introduced (“You will be shown terms that Ear, Nose and Throat (ENT) surgeons may use when communicating the risks associated with ENT surgery”). After providing demographic details, study instructions were titrated to participants. The first set of instructions primarily pertained to the meaning of lower bound, upper bound and best estimates. Participants read: “In this study, we are interested in your understanding of verbal probability expressions. **Verbal probability expressions** are terms that describe the likelihood of the occurrence of an event, such as below” [emphasis in original]. Participants were then presented with the list of VPEs to be used in the study (see Figure 2), before being told that VPEs “often convey a range of probabilities, and consequently have a lower and an upper bound, and a best estimate.” The meanings of these terms were provided to participants, for example, “a **lower bound estimate** is the *lowest* numerical probability that you think the term can represent”; “A **best estimate** is the numerical probability that you think *best matches* the term.” Participants were informed that they were required to provide an estimate for each of these properties of each VPE.

The second set of instructions was presented below the first set upon clicking ‘next’ and (as for the surgeons) introduced the different numerical formats that participants could use for their answers. In addition, lay participants were provided with the example of four ways the probability of a coin toss landing heads could be described (50%, 1 in 2, 0.5, 1:1⁸). On the following page, the ENT surgery context was reinforced:

⁸ This example removed the need to ask lay participants about their understanding of a 1:2 ratio.

“Imagine you are a patient who is scheduled for surgery concerning the ear, nose and/or throat (ENT). Your ENT surgeon is required to inform you about the possible risks associated with the surgery. To do so, the surgeon might use **verbal probability expressions** to describe the probability of you experiencing a risk associated with the surgery.

You will be shown a table containing some terms a surgeon may use. These terms often convey a range of numerical probabilities.”

Below this text, participants received the same task instructions as surgeons (see above), but the table containing the VPEs was only presented (below these instructions) upon clicking ‘next.’ The one change to this table, from the one shown to surgeons, was emphasising in the column heading (in red text) that lower bound estimates ‘must be less than or equal to the best estimate’ and upper bound estimates ‘must be greater than or equal to the best estimate’ (although it was possible for participants to enter data that violated this condition). After providing an interpretation for all VPEs, participants were debriefed, thanked and redirected to Prolific for payment.

Results

Because participants were able to provide responses in any numerical format, responses where an integer was provided with no units were ambiguous (e.g., should it be interpreted as a percentage or a 1 in X response?). Such responses, as well as range responses, were

consequently unusable and excluded from analyses. In total, 447 responses (14%) were missing or unusable (280 [21%] from the surgeons; 167 [9%] from the lay participants⁹).

Preliminary analyses

To provide an indication of the quality of the data collected, we counted the number of participants displaying at least one violation of the requirement that upper bound estimates should be no less than best estimates, which in turn should be no less than lower bound estimates. 22/104 participants displayed at least one such inequality violation (10 in the surgeon sample, 12 in the laypeople sample; a difference that was not significant, $\chi^2(1) = 0.11, p = .74$). Participants violating these inequalities were not excluded from the analyses below. All reported significance patterns for Study 2 are unchanged if such participants are excluded from these analyses, with the exception of one interaction (see Footnote 10).

Best estimates

Two laypeople provided the same best estimate (50%) for each VPE. Given that their range estimates differed between the expressions, we saw no reason to exclude their data, whilst noting that all reported significance patterns are unchanged if they are excluded.

Figure 2 displays the ‘best estimates’ of the VPEs across both surgeons and laypeople. A clear pattern is apparent, whereby lay participants provided higher probability estimates than surgeons across all ten VPEs. This pattern is confirmed by a main effect of participant group in a 10×2 ANOVA (including only those participants with no missing data - 21 surgeons; 53 laypeople), $F(1, 72) = 31.9, p < .001, \eta_p^2 = .31$. This was in addition to a main effect of VPE, $F(2.9, 211.6) = 119.20, p < .001, \eta_p^2 = .62$. The interaction term was non-significant, $F(2.9, 211.6) = 1.67, p = .18, \eta_p^2 = .02$ (Greenhouse-Geisser corrections applied to repeated

⁹ The difference in missing responses is most likely due to different expectations in the two samples. As they received no remuneration, surgeons were probably less likely to feel any obligation to complete all the questions, especially as they were included within a longer questionnaire.

measures effects due to a violation of the sphericity of variance assumption)¹⁰. As a check of the robustness of the main result, Mann-Whitney tests were performed to compare the interpretations of surgeons and laypeople for each individual VPE. As illustrated in Figure 2, these tests demonstrated the robustness of the main effect of participant group, which held in non-parametric analyses of each individual VPE (both including all usable datapoints for the full sample [44 surgeons; 60 laypeople] and in an analysis where individual participants were initially excluded if they failed to provide usable ‘best estimates’ for at least half the VPEs [seven surgeons; five laypeople]).

Range of interpretations

One lay participant (ID = 68) was excluded from these analyses as all their upper bound estimates were lower than their lower bound estimates. In an overall analysis, including only those participants with no missing data (21 surgeons; 53 laypeople), no main effect of participant group was observed, $F(1, 72) = 0.36, p = .55, \eta_p^2 = .005$. In addition to a main effect of VPE, $F(3.7, 268.0) = 18.94, p < .001, \eta_p^2 = .21$, an interaction between VPE and participant group was observed, $F(3.7, 268.0) = 4.95, p = .001, \eta_p^2 = .06$. From Figure 3 and Table 3, we can see that laypeople provided larger ranges than surgeons for VPEs denoting lower probabilities, with this tendency attenuated for higher probabilities, and reversed for ‘likely’.

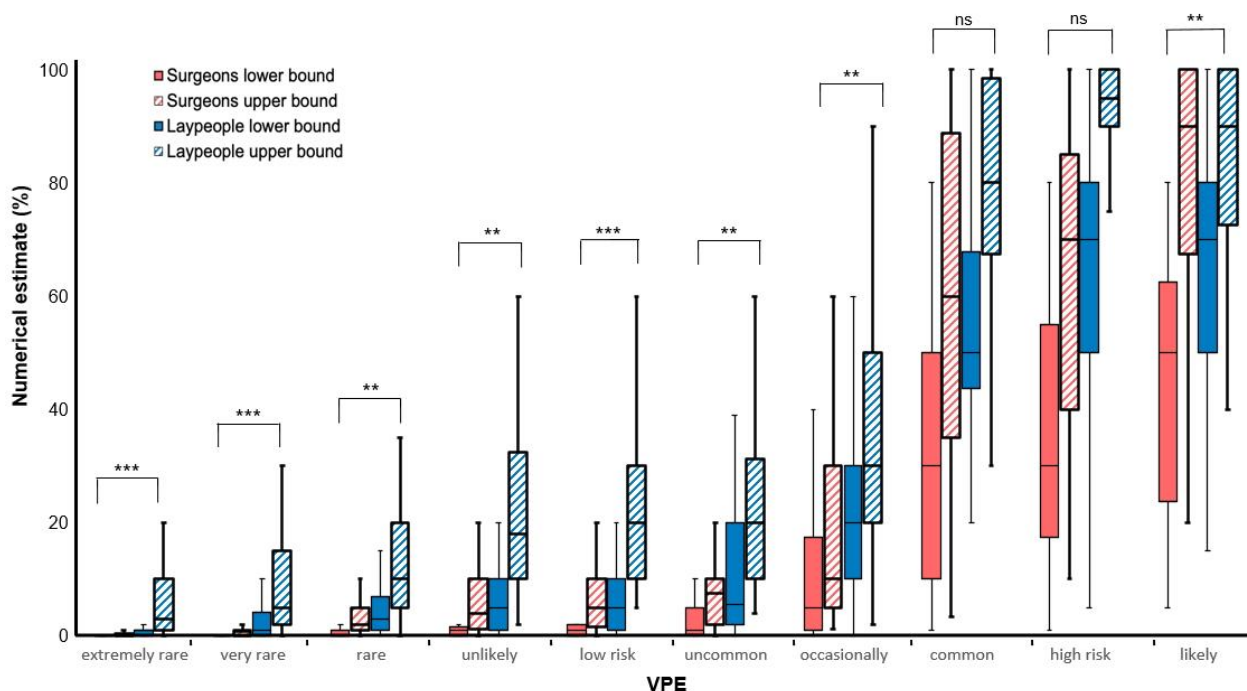
A closer inspection of Figure 3 suggests that the greater ranges provided by laypeople for VPEs denoting lower probabilities result from laypeople’s provision of higher upper bound estimates than surgeons – with surgeons providing lower lower-bound estimates for all VPEs. These main effects of participant group were significant for both lower bound, $F(1, 72) = 27.52, p < .001, \eta_p^2 = .28$, and upper bound estimates, $F(1, 72) = 19.74, p < .001, \eta_p^2 = .22$, in a 10

¹⁰ The interaction term was significant in the analysis that excluded participants who violated at least one of the inequality conditions, $F(4.5, 242.3) = 3.74, p < .01, \eta_p^2 = .07$, although the direction of the effect of participant group was consistent across all VPEs.

× 2 ANOVA. For the lower bound estimates, this main effect was qualified by an interaction with VPE, $F(4.2, 305.3) = 4.33, p = .002, \eta_p^2 = .06$, which was not observed for the upper bound estimates, $F(3.6, 258.2) = 1.25, p = .29, \eta_p^2 = .02$. Mann-Whitney tests (both including all usable datapoints for the full sample [44 surgeons; 59 laypeople] and where individual participants were initially excluded if they failed to provide usable ‘best estimates’ for at least half the VPEs [seven surgeons; five laypeople]) showed that surgeons’ lower bound estimates were significantly lower than laypeople’s across all VPEs, whilst their higher bound estimates were significantly lower across all VPEs, with the exception of ‘likely’.

Figure 3

Lower and upper bound estimates for ten VPEs from both surgeons and laypeople.



Note. The solid line within the box represents the median, the limits of the box represent the interquartile range, and the whiskers represent the range. Significance levels represent differences in the size of the ranges (differences between upper and lower bounds) between surgeons and laypeople. * $p < .05$; ** $p < .01$; *** $p < .001$ [by Mann-Whitney U tests]¹¹.

Table 3

Median (IQR) interpretation ranges (upper bound [%] MINUS lower bound [%]) provided by surgeons and laypeople in Study 2.

VPE	Surgeons	Laypeople
Extremely rare	0.1 (0.5)	2.0 (6.3)
Very rare	0.5 (0.9)	4.0 (8.7)
Rare	1.5 (3.1)	5.0 (8.0)

¹¹ These results were the same whether the analysis included or excluded participants who failed to provide usable lower and upper bound estimates for at least half the VPEs (seven surgeons; five laypeople).

Unlikely	3.0 (7.0)	10.0 (16.0)
Low risk	4.0 (8.1)	9.5 (14.0)
Uncommon	4.8 (9.0)	10.0 (13.8)
Occasionally	5.0 (6.3)	10.0 (11.0)
Common	20.0 (30.0)	20.0 (30.0)
High risk	20.0 (30.0)	20.0 (20.0)
Likely	30.0 (20.0)	15.0 (15.0)

Study 3

The reliable differences in VPE interpretations between surgeons and laypeople observed in Study 2 suggests potential for misunderstanding the likelihood of surgical risks, thus undermining the notion of fully informed consent. In Study 2, laypeople typically interpreted VPEs as denoting *higher* numerical probabilities than did surgeons. If this result is replicated in real surgical encounters, laypeople (i.e., patients) might refuse more surgeries than they might otherwise, due to an exaggerated perception of the likelihood of adverse outcomes – adverse outcomes being communicated via VPEs (Study 1) and interpreted as denoting higher probabilities than intended by surgeons (Study 2).

Study 3 was designed to provide a first step to testing a more appropriate lexicon for the communication of surgical risk. Specifically, we asked what VPEs laypeople intuitively generated to match surgeons' numerical interpretations from Study 2.

Method

Participants

Forty-eight female and 52 male participants aged between 18 and 71 (median = 30) were recruited online through Prolific.co and received £0.65 (average payment rate £8.74 per hour). The sample size of 100 was determined before data collection commenced and was deemed sufficient to identify clear agreements on appropriate VPEs. Ninety-five participants were native English speakers.

Design, Materials and Procedure

Participants were presented with ten numerical probabilities (presentation order randomised between participants), which closely approximated surgeons' 'best estimates' of frequently used VPEs (from Study 2: 0.01%, 0.1%, 0.2%, 0.5%, 2%, 5%, 10%, 20%, 50%, 60%). These probabilities were presented in numerical (percentage and '1 in X'¹² - apart from 60%, which was only presented as a percentage) and visual formats (the visual format followed Dhami, 2018; see Figure 4).

Participants provided informed consent, where the ENT context was specified (in addition to a title of "Risk communication about ENT surgery", participants were informed that they would "be shown numerical probabilities that Ear, Nose and Throat (ENT) surgeons may have to communicate to patients when communicating risks associated with ENT surgery"). After providing demographic details, participants read introductory text outlining the background of the study, which again reinforced the context of ENT surgical risk communication and our aims (whether there is a set of probability words "for which people share a degree of agreement about the likelihood they describe"). On the next page, participants entered ten VPEs (see Figure 4) before subsequently being thanked and debriefed. Participants received a single exemplar VPE to help in the description of the task (following Dhami, 2018). To control for the influence of this example VPE, half participants received 'likely'¹³ (as in Dhami, 2018) and half received 'unlikely'.

¹² Although '1 in X' formats have been shown to be poorly understood (e.g., Cuite et al., 2008; Sirota & Juanchich, 2019; Sirota, Juanchich & Bonnefon, 2018), they are the most common numerical format medical professionals report using (Sirota, Juanchich, Petrova, et al., 2018).



¹³ Due to a programming error, the *number* 60% was not presented to these participants (its visual representation still was).

Figure 4

How the numerical probabilities were presented to participants in the study. Ten probabilities were presented on this page.

In this study, we would like you to imagine being a patient scheduled for surgery concerning the ear, nose and/or throat (ENT). We are interested in what words (verbal probability expressions, e.g., 'likely') you would like a surgeon to use when communicating risks of the following approximate likelihoods?

(Below is a list of numerical probabilities with corresponding graphics. In each graphic, imagine there is a very thin, precise spinner. The probability is equal to the chance of the spinner landing in the blue area.)

	Verbal probability expressions
<p>1 in 2 (50%)</p> 	<input type="text"/>
<p>1 in 20 (5%)</p> 	<input type="text"/>

Results

Five participants were excluded from further analysis as they only reported numerical probabilities. One additional participant was excluded for responding ‘likely’ to all numerical probabilities. The presented results are thus based on 94 participants. Figure 5 demonstrates the considerable number of different VPEs generated, across the 94 participants, for each numerical probability.¹⁴ In addition, those VPEs generated most frequently overall (namely ‘unlikely’, ‘very unlikely’, ‘extremely unlikely’, ‘likely’, ‘highly unlikely’, ‘small chance’) were each generated to represent a variety of numerical probabilities (Figure 6). Results such

¹⁴ Note that these figures should be considered approximate, as it depends on how those phrases are differentiated (e.g., ‘near impossible’, ‘near on impossible’ and ‘nearly impossible’ are treated as the same phrase, whilst ‘near zero’ is considered as a unique one).

as the equal frequency of 'unlikely' and 'likely' to represent a 20% chance clearly illustrate the heterogeneity in perceptions of appropriate VPEs in the current domain. Figure 7 more clearly illustrates this spread of VPE generation by normalising according to the total number of uses of each individual VPE. Whilst these results would suggest that developing an evidence-based lexicon for the verbal communication of the entire range of surgical risks will be a challenge, the most striking result of Study 3 is shown in Figure 8. Only 14 of the 94 participants provided a different VPE for all ten numerical probabilities. This suggests that an appropriate lexicon likely does not exist to communicate the required range of surgical risk probabilities. It was also noteworthy that nine participants felt it necessary to use at least one number in their interpretations (not including responses for 50%), which were not included in Figure 8. This potentially further highlights the difficulty associated with representing the required range of probabilities with VPEs. Appendix B provides a breakdown of the VPEs most frequently generated for each numerical probability.

Figure 5

Number of different VPEs generated across the 94 participants for each numerical probability

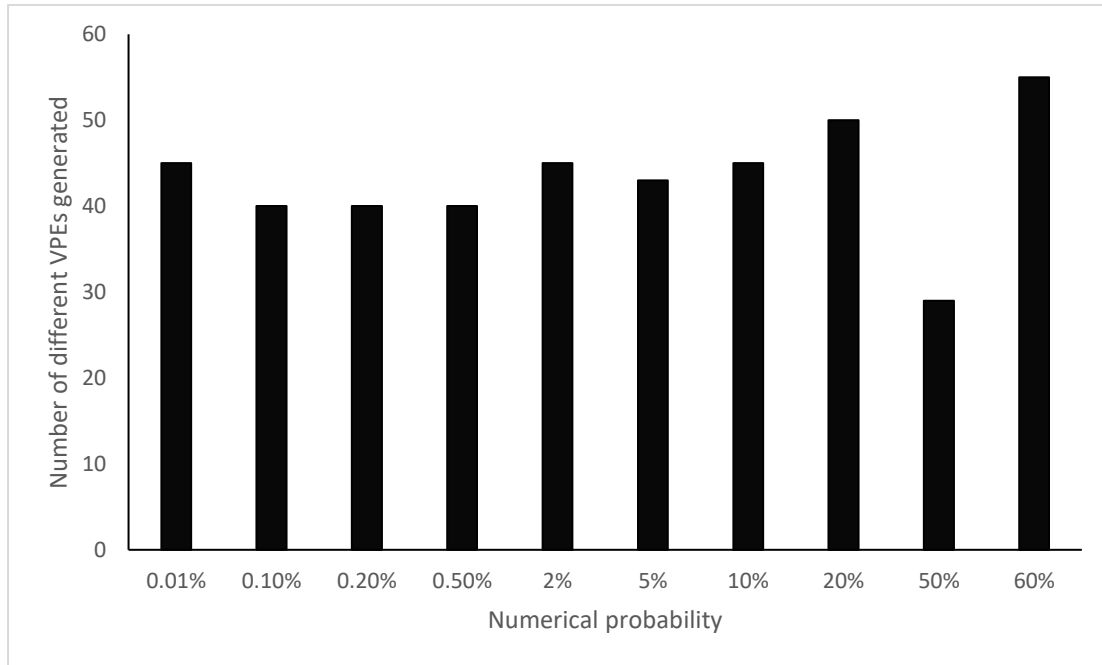


Figure 6

Number of times each of the six most popular VPEs were generated for each numerical probability

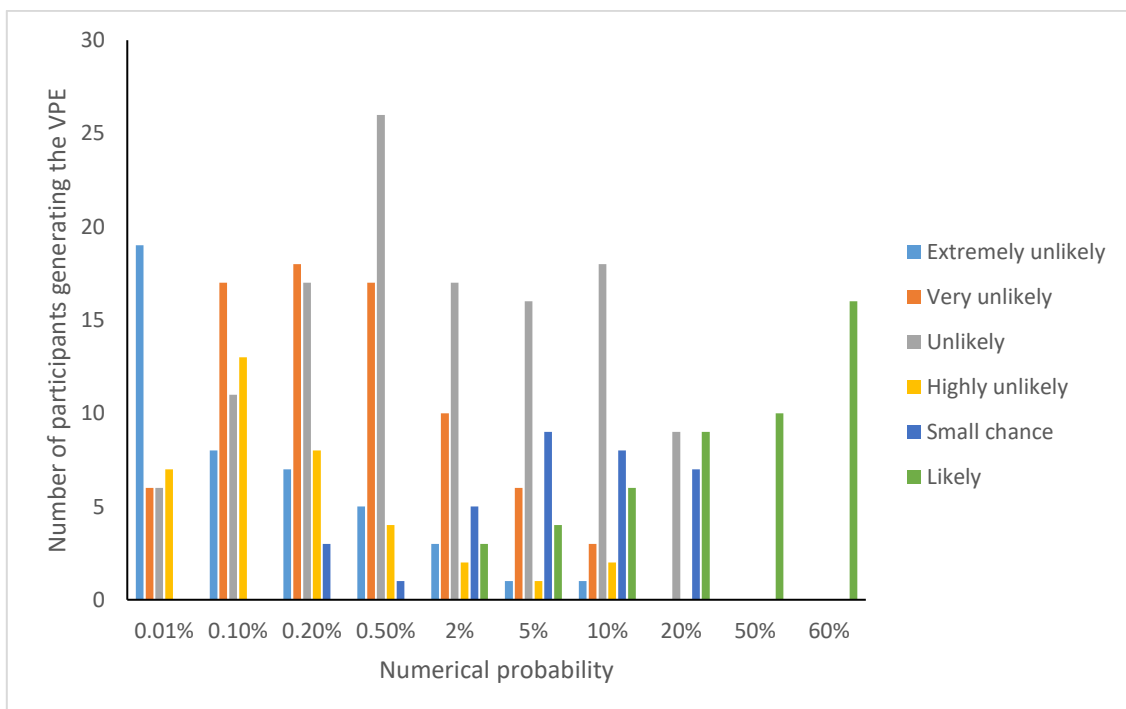


Figure 7

Distribution of VPE generation across the numerical probabilities, normalised as a percentage of the number of times each VPE was generated across the entire dataset

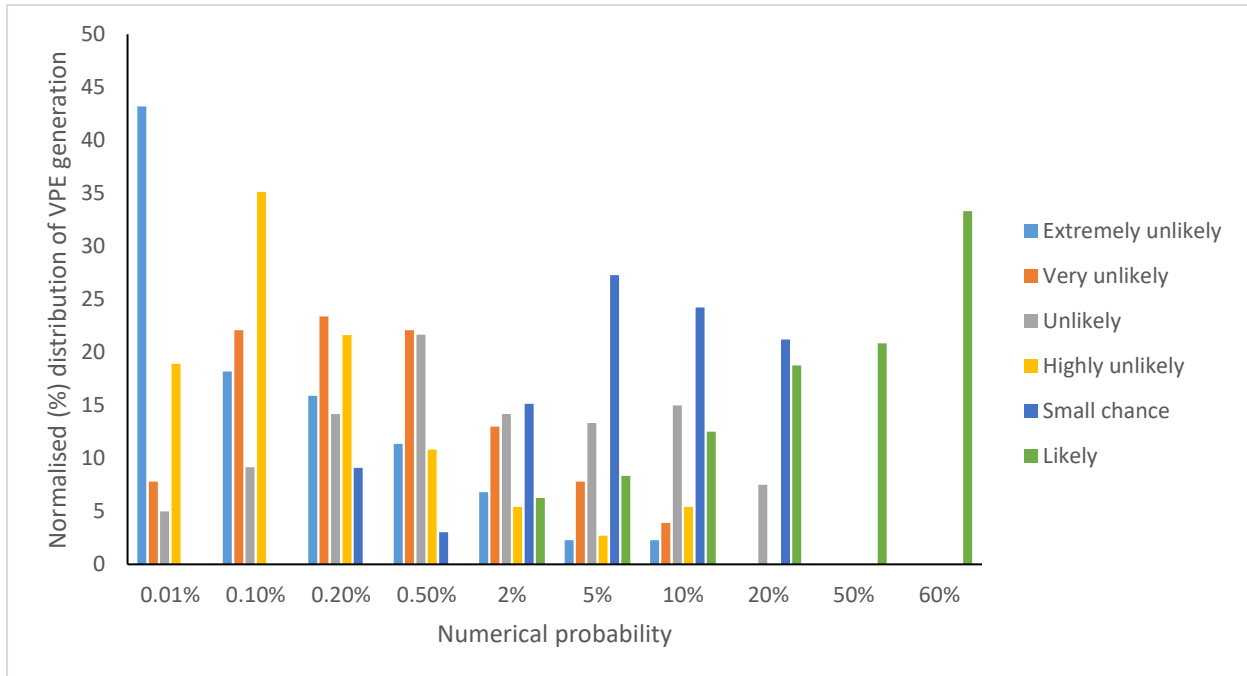
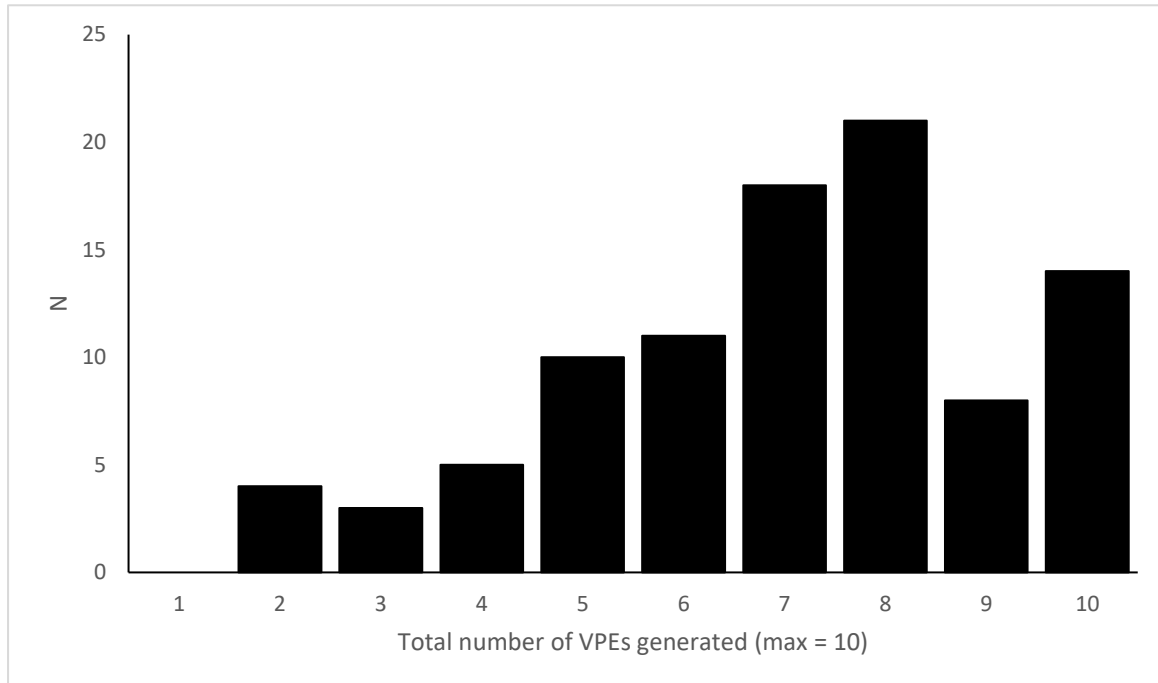


Figure 8

Number of different VPEs generated by participants to represent the 10 numerical probabilities (median = 7)

**Discussion**

The results of Study 3 demonstrated considerable heterogeneity in laypeople's chosen VPEs for ten numerical probabilities. The highest proportion of participants generating the same term in response to a numerical probability was 26 participants (28%) entering 'unlikely' for a 0.5% risk (see Figure 6 & Appendix B). In addition, only 15% of participants were able to generate a unique VPE for each of the ten probability levels. Whilst further analyses could investigate the effect of probability level on VPE heterogeneity, the message seems clear enough: VPEs are heterogeneous and non-specific. VPEs' inability to provide sufficient granularity (what Friedman et al., 2018, would refer to as 'probabilistic precision') to differentiate sizeably different risk levels ('unlikely' was generated by 18 participants for a 10% chance and 17 participants [3 of them the same participants] for a 0.1% chance – where the latter represents a risk level 100 times greater than the former) seems a strong argument for

the use of numbers in surgical risk communication. Moreover, this result suggests that research attention should turn to identifying optimal approaches to numerical risk communication and encouraging the adoption of such communication formats by surgeons.

General Discussion

Three studies were undertaken to test the potential for an ‘illusion of communication’ between ENT surgeons and laypeople (i.e., potential patients), and subsequently to investigate means for improving communication through an evidence-based standardised lexicon. Studies 1 and 2 provided the first conceptual replications of previous results observed in medical and other domains in the specific context of ENT surgery risk (and are among only a handful of studies investigating *surgical* risk generally). Study 1 demonstrated considerable heterogeneity in the VPEs currently used by ENT surgeons to communicate surgical risks. Study 2 demonstrated heterogeneity in how Study 1’s most frequently cited VPEs were interpreted, amongst both surgeons and laypeople. Moreover, laypeople consistently provided higher numerical interpretations of these VPEs than did surgeons. Such a result may be consequential. Overestimating surgical risk may lead a patient to turn down a beneficial surgical procedure that they would have accepted had they had a correct understanding of the risks involved. Study 3 was intended to be the first step in generating an evidence-based lexicon to improve verbal communications of surgical risk through identifying a set of VPEs that might be better suited to communicate the risk levels that surgeons need to convey to laypeople. Study 3 revealed considerable heterogeneity in people’s choices of VPEs. Such a result is not necessarily unexpected, and represents a challenge rather than an insurmountable barrier to the development of an appropriate lexicon (see e.g., Ho et al., 2015). However, the inability of most participants (85%) to generate a unique VPE for each level of risk (Study 3) suggests that attempts to develop an evidence-

based lexicon for risk communication about ENT surgery with VPEs may be doomed to failure.

It is worth noting that laypeople's inability to provide a unique VPE for each risk level might reflect the lack of a need to differentiate these risk levels for effective decision making. That is, sensitivity analyses may subsequently suggest that the difference between a 1 in 1,000 and a 1 in 10,000 risk will almost never make a difference to the decision a patient makes (or should make), even if they have a precise understanding of these probabilities. Consequently, an appropriate lexicon might be definable to delineate, for example, five relevant levels of risk (instead of the 10 levels we aimed for in the present study). Such an approach would raise the question (for ethicists and policy makers) of the degree to which *informed* consent is obtained in such situations. Our data suggest, however, that VPEs are ill-suited to communicate all of the risk levels that accompany ENT surgical procedures. Rather, it seems the *only* way to differentiate between all the levels of likelihood for severe consequences that can be associated with surgery is with a numerical representation.

Relatedly, one approach to improving risk communication advocated by some previous researchers is to supplement VPEs with numerical translations (e.g., Berry & Hochhauser, 2006; Budescu et al., 2009, 2012, 2014; Harris et al., 2017; Visschers et al., 2009; Wintle et al., 2019; Witteman & Renooij, 2003). Other recent research, however, gives reason for caution here. Jenkins et al. (2018, 2019) and Juanchich and Sirota (2017) found that problematic interpretations and consequences of VPE communications persisted even when subsequently supplemented with a numerical range translation, although these issues were partly ameliorated where the numerical probability was presented *before* the VPE (Jenkins et al., 2018, 2019). Mandel and Irwin (2021) additionally reported no communicational benefit (in terms of agreement with the NATO intelligence lexicon) from

the inclusion of VPEs with numerical ranges. From the recent research, it is unclear why VPEs should be included in communications at all.

Research in the medical domain has reported understanding of side effect risk to be unaffected (Moraes & da Silva Dal Pizzol, 2018), or even harmed (Knapp et al., 2015), where solely numerical communications were supplemented with verbal expressions (e.g., “Common: may affect up to 1 in 10 people”). In the one instance where understanding was improved, Sinayev et al. (2015) presented their combined format with the numerical percentages *before* the verbal label (and in point form). Thus, whilst there are documented problems associated with combined communication formats, these results - coupled with those of Jenkins et al. (2018, 2019) - suggest that research might profitably further evaluate the benefits of combined formats where the numerical information is presented before the verbal label. Such research is likely to be especially valuable if disposing of verbal terms completely proves to be a step-too-far for risk communicators.

One beneficial function of verbal terms is to contextualise numerical information. Including *evaluative* labels (e.g., ‘borderline’, ‘fair’) with numerical information has been shown, for example, to improve gist memory for screen test results (Morrow et al., 2018), and facilitate the use of relevant numerical information in judgments of hospitals (Peters et al., 2009). Where evaluative labels are required to help a patient contextualise a risk, it is important that the communicative purpose of the label is made clear to patients¹⁵, such that it should not be misunderstood as representing a statistical risk, but as representing an *evaluation* of the underlying risk. Moreover, there is a fine line between helping to contextualise and evaluate a risk, and providing an overly ‘directive’ communication (c.f. Zikmund-Fisher et al., 2007; see also Dieckmann et al., 2012). Consequently, further research

¹⁵ Such a recommendation is in line with Collins & Mandel (2019). Collins and Mandel recognised the benefit of VPEs in communicating *recommendations*, but argued that such recommendations should be explicit and where merely information is to be communicated, VPEs (with their implicit recommendations) should be avoided.

might investigate the question of how such ‘contextualising’ verbal expressions can best be presented to avoid such ethical concerns as well as those interpretation errors highlighted in past research.

As an alternative to VPEs, we also encourage researchers to further explore the benefits of supplementing numerical communications with visual aids, which have been shown to increase understanding, especially for individuals with lower numeracy skills (for reviews see e.g., Garcia-Retamero & Cokeley, 2013, 2017; Trevena et al., 2021). The fact that GPs report that they would use visual aids more frequently (for reporting cancer screening risks) when patients were described as being low in numeracy is an encouraging result in this context (Petrova et al., 2018).

Regardless of whether presented alone, or supplemented with words or visual aids, the precise format of a numerical risk communication should be chosen with care and extensively pilot tested before use (Bonner et al., 2021; Trevena et al., 2013, 2021). Bonner et al.’s (2021; see also Trevena et al., 2013, 2021) review of the literature on the effectiveness of different numerical representations should serve as a starting point here. As an example, they explicitly argue against ever using “1-in-X” formats, given people’s communication difficulties with them (e.g., Cuite et al., 2008; Sirota, Juanchich, & Bonnefon, 2018). An alternative approach (see e.g., Grimes & Snively, 1999; Lipkus, 2007; Trevena et al., 2021) would be to maintain a common (high) denominator, which would likely enable a clearer comparison of relative risks (e.g., - in the case of the probabilities used in the present Study 3: 1 in 10,000; 10 in 10,000; 20 in 10,000; 50 in 10,000; 200 in 10,000; 500 in 10,000; 1000 in 10,000; 2000 in 10,000; 5000 in 10,000; 6000 in 10,000). It is beyond the scope of the current manuscript to identify the *best* format of numerical communication. We hope, however, that the current results provide empirical incentive for such research to be further developed.

Despite extant evidence-based recommendations to favour numerical communication formats (Bonner et al., 2021; Trevena et al., 2013), Study 1 demonstrated that the majority of surgeons reported using words more often than numbers for risk communication. An open question that has received scant empirical attention is how can communicators be encouraged to use numerical communication formats. We are aware of only one published report investigating this issue, and this is from the intelligence domain (Barnes, 2016). The identified approaches are not readily transferable to the surgical domain for two reasons: 1) The Intelligence analysts were typically working with single-event subjective probabilities, whilst the probabilistic assessments that surgeons make are typically frequentist in nature (based on data from a population of previous patients). Consequently, training in the concept and application of single-event subjective probabilities, for example, will be of less relevance for surgeons; 2) An approximate ‘integer-in-10’ format for making numerical estimates was proposed to overcome concerns about the specificity implied by numerical estimates. This is unlikely to be appropriate for surgical risk communications which often entail differentiating between very small probabilities (as outlined throughout the current paper; see also Wiles et al., 2020). Whilst the lessons learned from Barnes (2016) are thus difficult to directly apply to the surgical domain, it is clear that encouraging the use of numerical communication formats is an important avenue for future research to explore. Such encouragement (and/or training) seems especially important for less numerate medical professionals who have been found to be the least likely to use such formats (Anderson et al., 2011; Petrova et al., 2018). One possibility to be explored, following from the revision of the scale according to analysts’ requirements (collapsing 2-3 and 7-8 out of 10; Barnes, 2016), might be to provide a non-linear scale of decisionally-relevant numerical probabilities for the surgical domain. Such a scale might enhance both understanding and uptake of numerical communication formats (see also Bonner et al., 2021). In essence, the numerical ranges in the EU’s prescribed lexicon

for communicating side effects (MHRA, 2015; see Table 2; see also Calman, 1996) provides such a scale (see also Paling, 2003; Woloshin et al., 2000, for additional possibilities). We maintain, however, that the scale should be evidence-based (within specific risk communication contexts) and additionally scrutinised by ethicists and policy makers to ensure it meets the requirements of informed consent.

Limitations

Our conclusions about surgeons' format preferences and VPE interpretations are based on a single sample of surgeons, with a similar level of experience (they were all Specialist Registrars [SpRs]). Although this might provide some limitation to the generalisability of these findings, SpRs are all trained to obtain informed consent from patients, both in medical school and in clinical practice, and are often responsible for the bulk of formal 'consenting' (e.g., explaining and completing consent forms). There is thus no reason to suspect that their assessment of the numerical risks of a given procedure would be any different from consultant surgeons.

Our studies also relied on self-reports from participants about how they typically communicate risk, and how they would interpret VPEs in the context of ENT surgical risks. Our confidence in the generalisability of the current results is strengthened by their consistency with those observed in previous research across a variety of domains, employing a range of methodologies and with a variety of professional participants. Moreover, Stheeman et al. (1993) observed the greatest interpersonal variance in professionals' interpretations for those terms that featured most frequently in a prominent dentistry (specifically dental radiography) textbook. Consequently, it is unlikely that the somewhat pessimistic conclusions about VPE use in the present article are a function of a mis-reporting of commonly used VPEs by the surgeons in Study 1. Nonetheless, an improved methodology might record real

consultations between surgeons and patients. This would have the added advantage of capturing the dialogical context of the risk communication (General Medical Council, 2008; Kunneman et al., 2015; see also Collins & Hahn, 2018, who call for greater attention to communicative context in the study of VPEs). Dialogue between patient and surgeon will be important, as patients can request additional information and clarification where required. Avoiding including potentially misleading VPEs in such dialogues would, however, still seem beneficial for effective surgical risk communication.

Conclusion and recommendations

The evidence presented here adds to a body of literature demonstrating the potential pitfalls of communicating with VPEs. Moreover, the research presented in Study 3 suggests that these pitfalls are an inherent property of VPEs, such that there may not be such a thing as an appropriate evidence-based lexicon for VPEs to communicate surgical risks. Dhimi and Mandel (2021; see also Dhimi et al., 2015; Irwin & Mandel, 2019) argue that the Intelligence community (and potentially others) should move towards numerical communications of probability, also citing a lack of granularity as one reason. Bonner et al. (2021) recommend numerical formats for conveying probabilities in patient decision aids, whilst Webster et al. (2017) suggest that Patient Information Leaflets should communicate side effect risk using numerical ranges rather than VPEs. We echo these sentiments. Whilst we can only make such an argument strongly in the specific case of ENT surgery (strictly, in ENT SpRs), we expect the case to be true across a broad array of risk domains; essentially, in any situation where individuals typically must differentiate between very small, but very different probabilities (e.g., 1 in 1,000 vs. 1 in 10,000). Such situations might arise in medical or non-medical contexts.

References

- Anderson, B. L., Obrecht, N. A., Chapman, G. B., Driscoll, D. A., & Schulkin, J. (2011). Physicians' communication of Down syndrome screening test results: The influence of physician numeracy. *Genetics in Medicine, 13*(8), 744-749.
- Barnes, A. (2016). Making intelligence analysis more intelligent: using numeric probabilities. *Intelligence and National Security, 31*(3), 327-344.
- Berry, D. C., & Hochhauser, M. (2006). Verbal labels can triple perceived risk in clinical trials. *Drug Information Journal, 40*(3), 249-257.
- Berry, D. C., Holden, W., & Bersellini, E. (2004). Interpretation of recommended risk terms: differences between doctors and lay people. *International Journal of Pharmacy Practice, 12*, 117-124.
- Berry, D. C., Knapp, P. R., & Raynor, T. (2002). Is 15 per cent very common? Informing people about the risks of medication side effects. *International Journal of Pharmacy Practice, 10*, 145-151.
- Berry, D. C., Raynor, D. K., & Knapp, P. (2003). Communicating risk of medication side effects: An empirical evaluation of EU recommended terminology. *Psychology, Health & Medicine, 8* (3), 251-263.
- Beyth-Marom, R. (1982). How probable is probable? A numerical translation of verbal probability expressions. *Journal of Forecasting, 1*, 257-269.
- Bonnefon, J. F., & Villejoubert, G. (2006). Tactful or doubtful? Expectations of politeness explains the severity bias in the interpretation of probability phrases. *Psychological Science, 17*, 747-751.
- Bonner, C., Trevena, L., Gaissmaier, W., Han, P. K. J., Okan, Y., Ozanne, E., Peters, E., Timmermans, D., & Zikmund-Fisher, B. J. (2021). Current best practice for

- presenting probabilities in decision aids: fundamental principles. *Medical Decision Making*, 41(7), 821-833.
- British Medical Association (2020). *Doctors' titles explained*. [Retrieved June 17th, 2021 from [Doctors' titles explained - Toolkit for doctors new to the UK - BMA](#)].
- Brun, W., & Teigen, K. H. (1988). Verbal probabilities: ambiguous, context-dependent, or both? *Organizational Behavior and Human Decision Processes*, 41, 390-404.
- Bryant, G. D., & Norman, G. R. (1980). Expressions of probability: words and numbers. *New England Journal of Medicine*, 302(7), 411–411.
- Budescu, D. V, Broomell, S. B., & Por, H. H. (2009). Improving communication of uncertainty in the reports of the intergovernmental panel on climate change. *Psychological Science*, 20(3), 299–308.
- Budescu, D. V, Por, H. H., & Broomell, S. B. (2012). Effective communication of uncertainty in the IPCC reports. *Climatic Change*, 113(2), 181–200.
- Budescu, D. V, Por, H. H., Broomell, S. B., & Smithson, M. (2014). The interpretation of IPCC probabilistic statements around the world. *Nature Climate Change*, 4(6), 508–512.
- Budescu, D. V., & Wallsten, T. S. (1990). Dyadic decisions with numerical and verbal probabilities. *Organizational Behavior and Human Decision Processes*, 46, 240-263.
- Budescu, D. V., Wallsten, T. S. (1995) Processing linguistic probabilities: general principles and empirical evidence. In J. R. Busemeyer, R. Hastie, & D. L. Medin (Eds.), *Psychology of Learning and Motivation: Advances in Research and Theory* (vol. 32: *Decision Making from a Cognitive Perspective*) (pp. 275-318). Academic.
- Calman, K. C. (1996). Cancer: science and society and the communication of risk. *British Medical Journal*, 313, 799-802.

College of Policing (n.d.), *Intelligence Management: Delivering effective analysis* [<https://www.app.college.police.uk/app-content/intelligence-management/analysis/delivering-effective-analysis/?highlight=%22PHIA%20PROBABILITY%20YARDSTICK%22?s=%5C%22PHIA+PROBABILITY+YARDSTICK%5C%22#communicating-probability>]. Retrieved August 14th, 2020.

Collins, P. J., & Hahn, U. (2018). Communicating and reasoning with verbal probability expressions. *Psychology of Learning and Motivation*, 69, 67-105.

Dhami, M. K., (2018). Towards an evidence-based approach to communicating uncertainty in intelligence analysis. *Intelligence and National Security*, 33(2), 257-272.

Dhami, M. K., & Mandel, D. R. (2021). Words or numbers? Communicating probability in intelligence analysis. *American Psychologist*, 76(3), 549-560.

Dhami, M. K., Mandel, D. R., Mellers, B. A., & Tetlock, P. E. (2015). Improving intelligence analysis with decision science. *Perspectives on Psychological Science*, 10, 753-757.

Dieckmann, N. F., Peters, E., Gregory, R., & Tusler, M. (2012). Making sense of uncertainty: advantages and disadvantages of providing an evaluative structure. *Journal of Risk Research*, 15(7), 717-735.

Erev, I., & Cohen, B. (1990). Verbal versus numerical probabilities: Efficiency, biases, and the preference paradox. *Organizational Behavior and Human Decision Processes*, 45, 1-18.

- Friedman, J. A., Baker, J. D., Mellers, B. A., Tetlock, P. E., & Zeckhauser, R. (2018). The value of precision in probability assessment: evidence from a large-scale geopolitical forecasting tournament. *International Studies Quarterly*, *62*, 410-422.
- Garcia-Retamero, R., & Cokely, E. T. (2013). Communicating health risks with visual aids. *Current Directions in Psychological Science*, *22*(5), 392-399.
- Garcia-Retamero, R., & Cokely, E. T. (2017). Designing visual aids that promote risk literacy: A systematic review of health research and evidence-based design heuristics. *Human Factors*, *59*(4), 582-627.
- General Medical Council (2008). *Consent: Patients and Doctors Making Decisions Together*. [Retrieved from www.gmc-uk.org/guidance August 14th, 2020].
- Grimes, D. A., & Snively, G. R. (1999). Patients' understanding of medical risks: implications for genetic counseling. *Obstetrics & Gynecology*, *93*(6), 910-914.
- Harris, A. J. L., & Corner, A. (2011). Communicating environmental risks: Clarifying the severity effect in interpretations of verbal probability expressions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 1571–1578
- Harris, A. J. L., Por, H-H., & Broomell, S. B. (2017). Anchoring climate change communications. *Climatic Change*, *140*, 387-398.
- Hart, A., Maxim, L., Siegrist, M., von Goetz, N., da Cruz, C., Merten, C., Mosbach-Schulz, O., Lahaniatis, M., Smith, A., & Hardy, A. (2019). Guidance on communication of uncertainty in scientific assessments. *EFSA Journal*, *17*(1): 5520. <https://doi.org/10.2903/j.efsa.2019.5520>.
- Henneman, L., Marteau, T. M., & Timmermans, D. R. M. (2008). Clinical geneticists' and genetic counselors' views on the communication of genetic risks: a qualitative study. *Patient Education and Counseling*, *73*, 42-49.

- Ho, E. H., Budescu, D. V., Dhimi, M. K., & Mandel, D. R. (2015). Improving the communication of uncertainty in climate science and intelligence analysis. *Behavioral Science and Policy*, 1(2), 43-55.
- Holtgraves, T., & Perdue, A. (2016). Politeness and the communication of uncertainty. *Cognition*, 154, 1-10.
- Irwin, D., & Mandel, D. R. (2019). Improving information evaluation for intelligence production. *Intelligence and National Security*, 34 (4), 503-525.
- Jenkins, S. C., Harris, A. J. L., & Lark, R. M. (2018). Understanding 'unlikely (20% likelihood)' or '20% likelihood (unlikely)' outcomes: The robustness of the extremity effect. *Journal of Behavioral Decision Making*, 31(4), 572-586.
- Jenkins, S. C., Harris, A. J. L., & Lark, R. M. (2019). When unlikely outcomes occur: The role of communication format in maintaining communicator credibility. *Journal of Risk Research*, 22(5), 537-554.
- Juanchich, M., & Sirota, M. (2017). How much will the sea level rise? Outcome selection and subjective probability in climate change predictions. *Journal of Experimental Psychology: Applied*, 23(4), 386-402.
- Juanchich, M., & Sirota, M. (2020). Most family physicians report communicating the risks of adverse drug reactions in words (vs. numbers). *Applied Cognitive Psychology*, 34, 526-534.
- Juanchich, M., Sirota, M., & Butler, C. M. (2012). The perceived functions of linguistic risk quantifiers and their effect on risk, negativity perception and decision making. *Organizational Behavior and Human Decision Making*, 118, 72-81
- Karelitz, T. M., & Budescu, D. V. (2004). You say "probable" and I say "likely": improving interpersonal communication with verbal probability phrases. *Journal of Experimental Psychology: Applied*, 10, 25-41.

- Knapp, P., Berry, D. C., & Raynor, D. K. (2001). Testing two methods of presenting side effect risk information about common medicines. *International Journal of Pharmacy Practice, 9*(suppl.), R6.
- Knapp, P., Gardner, P. H., & Woolf, E. (2015). Combined verbal and numerical expressions increase perceived risk of medicine side-effects: A randomized controlled trial of EMA recommendations. *Health Expectations, 19*, 264-274.
- Kunneman, M., Stiggelbout, A. M., Marijnen, C. A. M., & Pieterse, A. H. (2015). Probabilities of benefit and harms of preoperative radiotherapy for rectal cancer: What do radiation oncologists tell and what do patients understand? *Patient Education and Counseling, 98*, 1092-1098.
- Kunneman, M., Stiggelbout, A. M., & Pieterse, A. H. (2020). Do clinicians convey what they intend? Lay interpretation of verbal risk labels used in decision encounters. *Patient Education and Counseling, 103*(2), 418-422.
- Lipkus, I. M. (2007). Numeric, verbal, and visual formats of conveying health risks: suggested best practices and future recommendations. *Medical Decision Making, 27*(5), 696-713.
- Mandel, D. R., & Irwin, D. (2021). Facilitating sender-receiver agreement in communicated probabilities: is it best to use words, numbers or both? *Judgment and Decision Making, 16*(2), 363-393.
- Mandel, D. R., Wallsten, T. S., & Budescu, D. V. (2021). Numerically bounded linguistic probability schemes are unlikely to communicate uncertainty effectively. *Earth's Future, 9*, e2020EF001526. <https://doi.org/10.1029/2020EF001526>
- Mastrandrea, M. D., Field, C. B., Stocker, T. F., Edenhofer, O., Ebi, K. L., Held, H., Kriegler, E., Mach, K. J., Matschoss, P. R., Plattner, G-K., Yohe, G. W., & Zwiers, F. W. (2010) *Guidance note for lead authors of the IPCC fifth assessment report on consistent treatment of uncertainties. IPCC cross-working group meeting on*

consistent treatment of uncertainties. [retrieved from https://www.ipcc.ch/site/assets/uploads/2017/08/AR5_Uncertainty_Guidance_Note.pdf August 14th, 2020].

Merz, J. F., Druzdzal, M. J., & Mazur, D. J. (1991). Verbal expressions of probability in informed consent litigation. *Medical Decision Making, 11*(4), 273-281.

MHRA: Medical and Healthcare Products Regulatory Agency Committee on Safety of Medicines (2005). *Always Read the Leaflet: Getting the Best Information with Every Medicine.* The Stationery Office.

Michie, S., Lester, K., Pinto, J., & Marteau, T. M. (2005). Communicating risk information in genetic counselling: An observational study. *Health Education & Behavior, 32*, 589-598.

Moraes, C. G., & da Silva Dal Pizzol, T. (2018). Effect of different formats for information on side effects regarding medicine users' understanding: A randomized controlled trial. *Patient Education and Counselling, 101*, 672-678.

Morrow, D., Azevedo, R. F. L., Garcia-Retamero, R., Hasegawa-Johnson, M., Huang, T., Schuh, W., Gu, K., & Zhang, Y. (2019). Contextualising numeric clinical test results for gist comprehension: implications for EHR patient portals. *Journal of Experimental Psychology: Applied, 25*(1), 41-61.

Nakao, M. A., & Axelrod, S. (1983). Numbers are better than words. *The American Journal of Medicine, 74*(6), 1061-1065.

Neuner-Jehle, S., Senn, O., Wegwarth, O., Rosemann, T., & Steurer, J. (2011). How do family physicians communicate about cardiovascular risk? Frequencies and determinants of different communication formats. *BMC Family Practice, 12*:15, 1-9.

- ODNI: Office of the Director of National Intelligence (2007). *Prospects for Iraq's Stability: A Challenging Road Ahead*. [Retrieved from <https://fas.org/irp/dni/iraq020207.pdf> August, 14th, 2020].
- Ohnishi, M., Fukui, T., Matsui, K., Hira, K., Shinozuka, M., Ezaki, H., Otaki, J., Kurokawa, W., Imura, H., Koyama, H., & Shimbo, T. (2002). Interpretation of and preference for probability expressions among Japanese patients and physicians. *Family Practice*, *19*(1), 7-11.
- Paling, J. (2003). Strategies to help patients understand risks. *British Medical Journal*, *327*, 745-748.
- Peters, E., Dieckmann, N. F., Västfjäll, D., Mertz, C. K., Slovic, P., & Hibbard, J. H. (2009). Bringing meaning to numbers: the impact of evaluative categories on decisions. *Journal of Experimental Psychology: Applied*, *15*(3), 213-227.
- Petrova, D., Kostopoulou, O., Delaney, B. C., Cokely, E. T., & Garcia-Retamero, R. (2018). Strengths and gaps in physicians' risk communication: A scenario study of the influence of numeracy on cancer screening communication. *Medical Decision Making*, *38*(3), 355-365.
- Pieterse, A. H., van Dulmen, S., van Dijk, S., Bensing, J. M., & Ausems, M. G. E. M. (2006). Risk communication in completed series of breast cancer genetic counseling visits. *Genetics in Medicine*, *8*, 688-696.
- Royal College of Surgeons of England (2020). *Surgery Career Paths*. [Retrieved July 9th, 2021 from <https://www.rcseng.ac.uk/careers-in-surgery/trainees/foundation-and-core-trainees/surgery-career-paths/>].
- Shaw, N. J., & Dear, P. R. F. (1990). How do parents of babies interpret qualitative expressions of probability? *Archives of Disease in Childhood*, *65*, 520-523.

- Sinayev, A., Peters, E., Tusler, M., & Fraenkel, L. (2015). Presenting numeric information with percentages and descriptive risk labels: a randomized trial. *Medical Decision Making*, 35, 937-947.
- Sirota, M., & Juanchich, M. (2019). Ratio format shapes health decisions: The practical significance of the “1-in-X” effect. *Medical Decision Making*, 39(1), 32-40.
- Sirota, M., Juanchich, M., & Bonnefon, J. F. (2018). “1-in-X” bias: “1-in-X” format causes overestimation of health-related risks. *Journal of Experimental Psychology: Applied*, 24(4), 431-439.
- Sirota, M., Juanchich, M., Petrova, D., Garcia-Retamero, R., Walasek, L., & Bhatia, S. (2018). Health professionals prefer to communicate risk-related numeric information using “1-in-X” ratios. *Medical Decision Making*, 38(3), 366-376.
- Stheeman, S. E., Mileman, P. A., van't Hof, M. A., & van der Stelt, P. F. (1993). Blind chance? An investigation into the perceived probabilities of phrases used in oral radiology for expressing chance. *Dentomaxillofacial Radiology*, 22, 135-139.
- Su, A. (2020). *Risk communication in ENT surgery – when severity is introduced, what are the changes and what do we think we know?* Unpublished undergraduate dissertation, University College London.
- Sutherland, H. J., Lockwood, G. A., Tritchler, D. L., Sem, F., Brooks, L., & Till, J. E. (1991). Communicating probabilistic information to cancer patients: is there ‘noise’ on the line? *Social Science & Medicine*, 32(6), 725-731.
- Theil, M. (2002). The role of translations of verbal into numerical probability expressions in risk management: a meta-analysis. *Journal of Risk Research*, 5, 177-186.
- Timmermans, D. (1994). The roles of experience and domain of expertise in using numerical and verbal probability terms in medical decisions. *Medical Decision Making*, 14(2), 146–156.

- Trevena, L. J., Bonner, C., Okan, Y., Peters, E., Gaissmaier, W., Han, P. K. J., Ozanne, E., Timmermans, D., & Zikmund-Fisher, B. J. (2021). Current challenges when using numbers in patient decision aids: advanced concepts. *Medical Decision Making, 41*(7), 834-847.
- Trevena, L. J., Zikmund-Fischer, B. J., Edwards, A., Gaissmaier, W., Galesic, M., Han, P. K. J., King, J., Lawson, M. L., Linder, S. K., Lipkus, I., Ozanne, E., Peters, E., Timmermans, D., & Woloshin, S. (2013). Presenting quantitative information about decision outcomes: a risk communication primer for patient decision aid developers. *Medical Informatics and Decision Making, 13* (Suppl. 2): S7.
- Visschers, V. H. M., Meertens, R. M., Passchier, W. W. F., & de Vries, N. N. K. (2009). Probability information in risk communication: A review of the research literature. *Risk Analysis, 29*, 267-287.
- Wallsten, T. S., Budescu, D. V., & Zwick, R. (1993). Comparing the calibration and coherence of numerical and verbal probability judgments. *Management Science, 39*, 176–190.
- Weber, E. U., & Hilton, D. J. (1990). Contextual effects in the interpretations of probability words: Perceived base rate and severity of events. *Journal of Experimental Psychology: Human Perception and Performance, 16*, 781–789.
- Webster, R. K., Weinman, J., & Rubin, G. J. (2017). People’s understanding of verbal risk descriptors in patient information leaflets: a cross-sectional national survey of 18- to 65- year olds in England. *Drug Safety, 40*, 743-754.
- Wiles, M. D., Duffy, A., & Neill, K. (2020). The numerical translation of verbal probability expressions by patients and clinicians in the context of peri-operative risk communication. *Anaesthesia, 75* (Suppl. 1), e39-e45.

- Wintle, B. C., Fraser, H., Wills, B. C., Nicholson, A. E., & Fidler, F. (2019). Verbal probabilities: *very likely* to be *somewhat* more confusing than numbers. *PLOS ONE*, *14*(4): e0213522.
- Witteman, C., Renooij, S. (2003) Evaluation of a verbal-numerical probability scale. *International Journal of Approximate Reasoning*, *33*, 117-131.
- Woloshin, S., Schwartz, L. M., Byram, S., Fischhoff, B., & Welch, G. (2000). A new scale for assessing perceptions of chance: A validation study. *Medical Decision Making*, *20*(3), 298-307.
- Ziegler, A., Hadlak, A., Mehlbeer, S., & König, I. R. (2013). Comprehension of the description of side effects in drug information leaflets: A survey of doctors, pharmacists and lawyers. *Deutsches Ärzteblatt International*, *110* (40), 669-673.
- Zikmund-Fisher, B. J., Fagerlin, A., Keeton, K., & Ubel, P. A. (2007). Does labeling prenatal screening test results as negative or positive affect a woman's responses? *American Journal of Obstetrics & Gynecology*, *197*, 528.e1-528.e6.

Appendix A

VPEs Generated by Surgeons in Study 1 (Listed in Descending Order of Frequency)

Expression	N	Expression	N	Expression	N
rare	33	apply to all operations	1	often	1
common	25	can happen but rare	1	probability	1
unlikely	19	chance	1	probable	1
uncommon	15	complication	1	rare but possible	1
very rare	14	death	1	rare but serious	1
low risk	9	definitely	1	rare but significant	1
extremely rare	8	extremely unlikely	1	relatively common	1
high risk	7	high potential	1	relatively high risk	1
likely	6	highly likely	1	relatively often	1
occasionally	5	highly unlikely	1	relatively rare	1
frequent	4	improbable	1	seen often	1
never seen	4	incredibly small risk	1	serious but rare risk	1
rarely	4	incredibly unlikely	1	small	1
sometimes	4	infrequent	1	small possibility	1
theoretical	4	life-changing	1	small proportion of patients	1
very unlikely	4	local rate	1	textbook risk	1
possibility	3	low possibility	1	unexpected problem	1
small chance	3	low potential	1	unlikely but possible	1
expected	2	may occur	1	unlikely but significant	1
high possibility	2	minimal	1	unlikely to happen	1
more common	2	moderate risk	1	unpredictable but unlikely	1

permanent	2	most	1	usual	1
possible	2	most common	1	usually	1
significant	2	most significant	1	very common	1
small risk	2	national rate	1	very often	1
temporary	2	not seen very often	1	very small chance	1
vanishingly rare	2	not usual	1	very uncommon	1
very low risk	2	not very likely	1	with great certainty	1

Appendix B. *The most frequently generated VPEs for each numerical probability.***0.01% chance**

19 participants entered 'extremely unlikely', with two participants entering 'extremely low' and three entering 'extremely low risk', 'extremely rarely' and 'extremely small.' The next most popular phrases were 'highly unlikely' (7), and 'very unlikely' and 'unlikely' (6). Four participants entered each of 'almost impossible', 'rare' and 'very rare'. The above entries constituted more than 50% of responses. Overall, approximately 45 phrases were entered by the 94 participants.

0.1% chance

17 participants entered 'very unlikely', with 13, 11 and 8 entering 'highly unlikely', 'unlikely' and 'extremely unlikely' respectively, thus making up more than 50% of responses. Overall, approximately 40 phrases were entered by the 94 participants, although two of these did suggest an error in interpretation ('almost certain', 'certain').

0.2% chance

18 and 17 participants entered 'very unlikely' and 'unlikely' respectively, with 8 and 7 respectively entering 'highly unlikely' and 'extremely unlikely', constituting more than 50% of responses. Overall, approximately 40 phrases were entered by the 94 participants.

0.5% chance

26 participants entered 'unlikely', with 17 participants entering 'very unlikely.' 5 participants entered 'extremely unlikely', constituting more than 50% of responses. Overall, approximately

40 phrases were entered by the 94 participants, with one reporting 'probably' and one 'significant' (likely reflecting a misunderstanding of the concept of statistical significance).

2% chance

17 participants entered 'unlikely', with 10 entering 'very unlikely', 5 entering 'small chance' and 4 entering each of 'possible' and 'slim chance.' This represents 43% of responses. Overall, approximately 45 phrases were entered by the 94 participants.

5% chance

16 participants entered 'unlikely', with 9 entering 'small chance' and 6 entering each of 'possible' and 'very unlikely.' This represents 39% of responses. Overall, approximately 43 phrases were entered by the 94 participants.

10% chance

18 participants entered 'unlikely', with 8 participants entering either 'small' or 'small chance', 6 participants entering 'likely' and 6 entering 'possible.' This represents 40% of responses. Overall, approximately 45 phrases were entered by the 94 participants.

20% chance

The modal response was less popular than for the risks analysed to date, with 9 participants entering 'likely' and 'unlikely' respectively. 'Possible', 'a chance', and 'quite unlikely' were entered by 6, 3, 3 participants respectively. Seven participants entered some variant of 'small chance' or 'small risk.' These responses constituted 39% of responses. A further 11 phrases were entered by two people. Interestingly, phrases such as 'very high probability', 'very risky', 'most likely' were also selected by participants. Approximately 50 phrases were entered by the 94 participants.

50% chance

'Likely' was entered by 10 participants, whilst the majority of other responses constituted phrases such as 'half', '50/50.' Four participants entered 'very likely' and two entered 'very

high,' with at least four more suggesting this reflected a high risk. Approximately 29 phrases were entered by the 94 participants.

60% chance

16 participants entered 'likely', with five entering 'highly likely' and 'very likely,' Four more participants entered each of 'possible' and 'quite likely.' These responses constituted 36% of responses. Approximately 55 phrases were entered by the 94 participants.