ORIGINAL RESEARCH

What are the Relevant Outcomes of the Periodic Health Examination? A Comparison of Citizens' and Experts' Ratings

Isolde Sommer [1]
Viktoria Titscher[1]
Monika Szelag[1]
Gerald Gartlehner[1,2]

[1]Department for Evidence-Based Medicine and Evaluation, Danube University Krems, Krems, Austria; [2]RTI International, Research Triangle Park, Raleigh, NC, USA

Purpose: Despite evidence from clinical guideline development that physicians and patients show discordance in what they consider important in outcome selection and prioritization, it is unclear to what extent outcome preferences are concordant between experts and citizens when it comes to the context of primary prevention. Therefore, the objective of this study was to assess whether expert judgments about the importance of beneficial and harmful outcomes differ from citizen preferences when considering intervention options for a periodic health examination (PHE) program.

Participants and Methods: We conducted an online survey using a modified Delphi approach. The target population for the survey consisted of citizens who had attended the PHE (n=18) and experts who made evidence-based recommendations (n=11). Citizens and experts assigned a score on a 9-point Likert scale for each outcome of 14 interventions. We analyzed the intragroup agreement based on Krippendorff's alpha and the intergroup agreement using the cube root product measure (CRPm). We further tested for significant differences between the groups using the Mann U-test.

Results: Agreements within the groups of citizens and experts varied across the interventions and tended to be poor ($\alpha \leq 0$ to 0.20) or fair ($\alpha = 0.21$ to 0.40), with three exceptions showing moderate agreement ($\alpha = 0.44$ to 0.55). The agreements between the citizens and experts across the interventions was fair (CRPm = 0.28) during the first Delphi rating round. The mean differences between the citizens and experts on the Likert scale ranged from 0.0 to 3.8 during the first rating round and from 0.0 to 3.3 during the second. Across interventions, the citizens rated the outcomes as more important than the experts did (p<0.01). Individual participants' ratings varied substantially.

Conclusion: Because experts generally underestimated the outcomes' importance to citizens, the involvement of citizens in guideline panels for preventive services is important.

Keywords: outcome ratings, guideline methodology, patient involvement, GRADE, health examination

Introduction

The past decades have seen a profound shift in modern health care systems from a paternalistic toward a patient-centered approach involving shared decision-making.[1] Although this approach is strongest within the individual patient–physician relationship, patients or citizens have an increasing opportunity to contribute, participate, and voice their perspectives during guideline development.[2]

Correspondence: Isolde Sommer
Department for Evidence-Based Medicine and Evaluation, Danube University Krems, Dr.-Karl-Dorrek-Straße 30, Krems 3500, Austria
Tel +43 (0)2732 893-2927
Fax +43 (0)2732 893-4910
Email isolde.sommer@donau-uni.ac.at

The importance of considering patients' or citizens' values and preferences during guideline development has been recognized by international institutions and groups[3,4] and is also reflected in the instruments that measure guideline quality. For example, the Appraisal of Guidelines for Research and Evaluation II (AGREE II) instrument to assess the quality of professional guidelines requires guideline developers to seek patients' views and preferences—the views of the target population.[5] However, guidance on how to engage patients in the process is still lacking. A recent systematic review on guidance documents for developing clinical guidelines from 56 institutions identified 40 (71.4%) that recommended the inclusion of patients or their representatives in the clinical guideline development process, with very few providing explicit guidance on how to implement this.[6]

To address this issue, Armstrong et al[7] proposed a framework outlining options for the methods and timing of patient engagement in 10 steps, ranging from identifying relevant guideline topics and evaluating the methods to impact of engagement in the guideline process. It also involves the contribution of patients in developing guideline questions, including the selection and prioritization of outcomes. One method to implement this is to conduct a survey wherein patients select and rate the importance of outcomes. The Grading of Recommendations Assessment, Development and Evaluation (GRADE) working group initially stressed the importance of focusing on outcomes that are important or critical for clinical decision-making from the viewpoint of those affected by the guideline.[8] In practice, however, the process of rating the relative importance of health outcomes is often performed by guideline panel members acting as proxies for patients or citizens.[8]

It is further known from research studies comparing outcome selection and prioritization that physicians and patients show discordance in what they consider important.[9–12] Recently, Gutmann et al[9] found notable differences between patients and their caregivers and a working group of nephrologists in the scope and focus of topics or outcomes of clinical practice guidelines for percutaneous renal biopsy. While the experts prioritized topics focused on clinical procedures and outcomes related to decreasing recovery time, minimizing the risk of complications, and maximizing sample yield, patients and caregivers were focused on self-management, psychosocial impact, education, patient–provider communication, and impact on family. A study by Demyttenaere et al[10] found that physicians attached more importance than patients did to depressive and anxious symptom severity, functional impairment, and some aspects of quality of life for the cure of depression. In contrast, patients emphasized the importance of positive affect.

Despite this evidence from clinical guideline development, it remains unclear to what extent outcome preferences are concordant between experts and citizens when it comes to the context of primary prevention. The primary preventive context differs from a curative context in that it targets citizens without the symptoms or awareness of having a disease. Citizens do not experience any burden of a disease, including discomfort, pain, symptoms, and its impact on their quality of life. They select and rate outcomes of a disease that does not affect them and that they aim to prevent.

The objective of our study was to assess whether expert judgments about the importance of beneficial and harmful health outcomes differ from citizen preferences when considering intervention options for the periodic health examination (PHE).

Methods

The present study was part of an evidence-based revision of the PHE program in the Austrian primary care setting. In Austria, statutory health insurances offer free annual health examinations to men and women aged 18 years or older.[13] This study was conducted in accordance with the Declaration of Helsinki.[14]

Design

To achieve our objective, we conducted a web-based survey including randomly selected citizens and multidisciplinary experts who were part of the guideline panel revising the PHE. The survey was open from December 2017 to April 2018 and used a modified Delphi approach with two rounds of rating the outcomes' importance.

Participants

The target population for our survey consisted of two groups: citizens who have attended the PHE and experts who make evidence-based recommendations about the PHE. Details about the citizen selection process have been previously published.[15] Briefly, regional health insurance funds identified a random sample of 3600 eligible citizens and sent postal invitation letters to participate in the study. Citizens could respond free of charge with a pre-addressed envelope, and 8.2% responded to the invitation

letters. Of those, we selected 30 persons based on an a priori developed sampling grid that included the following categories: age, gender, place of residence, education, and migration background. Before participating in the survey, citizens took part in focus groups that determined their expectations and experiences of the PHE and helped select outcomes they deemed relevant to deciding whether to participate in a PHE. The sample size of the focus groups was primarily guided by the heterogeneity of the citizens and data saturation. Ethical approval was granted by the Danube University Research Ethics committee (EK GZ 16/2015–2018). We sought informed consent from the participants in the form of signatures before the start of the focus group and implied consent from the participants of the survey. The Danube University Research Ethics committee approved both forms of consent.

Experts who participated in the survey were members of a multidisciplinary guideline panel for the revision of the Austrian PHE. We chose the panel members to achieve a diverse representation of backgrounds and expertise, resulting in an expert panel of 11 members who had education and experience in public health, general medicine, internal medicine, and evidence-based methods, or who were patient representatives.

Identifying Relevant Outcomes

In a first step, we collected potentially relevant outcomes for the PHE. We derived these outcomes from themes that emerged during the qualitative analysis of the data of the aforementioned focus groups comprised of citizens who had attended PHEs. These outcomes did not refer to specific screening and counseling interventions but rather to the PHE as a program (eg, improved quality of life or decreased risk of mortality). In a second step, to arrive at more specific outcomes for each intervention of interest, we searched the literature for systematic reviews and complemented the lists with outcomes identified in the relevant scientific literature. Finally, we asked the guideline panel members to identify additional relevant outcomes from their perspectives. To the experts we emphasized the importance of the lists including outcomes of both the benefits and harms of screening and counseling interventions. The list of outcomes for each intervention is displayed in the figures in the Supplementary material.

Web-Based Survey

We developed a web-based survey using the SurveyMonkey, Inc. software (San Mateo, California,

USA). In this survey, we asked citizens and experts to rate the outcomes' importance. In other words, we asked them to select and rate which outcomes they deem the most important when they must decide for or against a specific intervention. This approach follows the one proposed by GRADE,[8] which recommends that guideline developers consider multiple outcomes, providing a balance between benefits and harms for their decision-making process. GRADE rates outcomes on a scale from 1–9 and classifies them into three categories according to their importance for decision-making (critical, important but not critical, of limited importance).

For each intervention, we created an individual page in the web-based survey, asking the following question: if you are to decide whether an adult should participate in this intervention, how important are the following desirable and undesirable consequences in making such a decision? Participants were then prompted to assign each outcome a score on a 9-point Likert scale (1 is not important, and 9 is critically important). We pilot-tested the survey on four persons and revised the final survey based on their feedback.

The citizens and experts anonymously rated the outcomes. To ensure that the citizens understood the task and outcomes, we asked an experienced scientific writer to adapt the text and outcomes to plain language for lay audiences. For example, instead of "overtreatment," we used the term "risk of being treated even though you are healthy." We applied a modified Delphi method[16] using two rounds of ratings, so participants had the opportunity to reconsider their first-round rating with the group results in mind.

Statistical Analysis

After the first round, we calculated the means and ranges on the 9-point Likert scale for all the outcome ratings and reported them back to the participants at a group level (citizen and expert panels together). We did not impute missing data. To achieve consensus, we used the means and ranges on the 9-point Likert scale for all the second-round outcome ratings (citizens and experts together). The seven highest ranked endpoints for each intervention informed decision-making on the PHE program's intervention options.

To assess whether experts would have been able to judge the importance of beneficial and harmful outcomes to citizens, we analyzed the intergroup agreement between the citizens and experts. Agreement measures are useful

means of assessing how closely two observers agree in their assessments. If the agreement is poor, the suitability of the expert group to rate patient-centered outcomes can be questioned.[17] We used the cube root of the product measure (CRPm) to analyze the intergroup agreement.[17] This measure calculates the intragroup agreement within each group and in the combined group based on Krippendorff's alpha, a measure of agreement among several raters, using the formula:

$$CRP_m = \sqrt[3]{Agr(A) * Agr(B) * Agr(AB)}$$

Agr(A) corresponds to the agreement within the citizen group of m1 raters (m1 = number of raters in group 1), *Agr(B)* to the agreement within the expert group of m2 raters (m2 = number of raters in group 2), and *Agr(AB)* to the agreement within a combined group of m1+m2 raters. Krippendorff's alpha reaches 1 when there is perfect agreement among the raters. An alpha of 0 indicates no agreement, and a negative alpha shows systematic disagreement that exceeds what can be expected by chance. The CRPm is interpreted similarly to Krippendorff's alpha: if the raters in each group completely agree, the CRPm = 1; if there is no agreement or disagreement among the raters then the CRPm $\leq$ 0. Several benchmark scales have been constructed to aid in the interpretation of agreement data. The most widely used in the medical field is the one proposed by Altman.[18] It considers values below 0.20 as a poor strength of agreement, from 0.21 to 0.40 as fair, from 0.41 to 0.60 as moderate, from 0.61 to 0.80 as good, and from 0.81 to 1 as a very good strength of agreement.

To explore whether the agreement measures would change with fewer categories, we recoded the data using the three GRADE categories critical (7–9 points), important but not critical (4–6 points), and of limited importance (1–3 points) instead of the 9-point Likert scale and reran the analyses. We further calculated the means of the points on the Likert scale for each outcome and the ranges of the mean differences and used a Mann–Whitney *U*-test to determine whether the experts scored significantly higher than the citizens. To assess the rating variability, we computed the mean difference and standard deviation of the points on the Likert scale between the 10th and the 90th percentile of both the citizens and experts for each intervention. We used R[19] to conduct the statistical analyses. The analyses were done after the first and second survey round.

Results

Eighteen citizens and 10 experts participated in at least one round of the survey. Depending on the health outcome, 11 to 14 (out of 30) citizens and 9 to 10 (out of 11) experts completed the ratings. Overall, the participants rated 153 outcomes for 14 interventions (see Supplementary material for number of participants and outcomes for each intervention).

The results showed a generally fair agreement in the importance of the outcomes within the individual citizen and expert groups as well as between the two groups. During the first rating round in which participants were not yet influenced by the group results, the overall agreement across all outcomes was α=0.24 among the citizens and α=0.36 among the experts (Table 1), both of which are considered fair agreements. Table 1 provides the agreement coefficients for each PHE intervention for the citizens and experts as individual groups as well as between the two groups. In 8 of 14 interventions, the experts showed a higher within-group agreement than the citizens regarding the outcomes' importance. The agreement between the citizens and experts across all outcomes was fair, with a CRPm=0.28. The agreements between the citizens and experts were highest for health outcomes of screening for glaucoma, hearing impairment, and hepatitis C (>0.40). In contrast, agreements about the importance of the health outcomes of counseling on physical activity and screening for anemia, vitamin D deficiency, and osteoporosis were the lowest (<0.10). For the outcomes of screening for anemia, the citizens and experts achieved no agreement at all. Recoding the data into the three GRADE categories generally did not increase the agreement levels.

Comparisons of the mean points on the Likert scale for the outcomes across all interventions show that the citizens tended to rate the outcomes as statistically significantly more important than the experts did (p<0.01 when testing 9 Likert scores and 3 GRADE categories, Mann–Whitney *U*-Test, Table 1). When looking at individual interventions, those with significantly higher outcome ratings among the citizens compared to the experts are screening for anemia (p<0.01), parodontitis (p<0.01), age-related vision impairment (p<0.05), vitamin D deficiency (p<0.01), and chronic kidney diseases and urinary tract infections (p<0.05), with all the remaining outcomes statistically significant when testing the 3 GRADE categories, apart from screening for age-related vision impairment

Table 1 Agreements and Mean Differences Among and Between the Citizens and Experts in the 1st Round

Intervention (Number of Outcomes)	Agreement of Citizens 9-Point Likert Scale† (Recoded Data)§	Agreement of Experts 9-Point Likert Scale† (Recoded Data)§	Agreement Between Citizens and Experts 9-Point Likert Scale‡ (Recoded Data)§	Range of Mean Differences Between Citizens and Experts in Points on the 9-Point Likert Scale	Mean difference (Standard Deviation) in Points on the Likert Scale Between the 10th and the 90th Percentile of Citizens	Mean Difference (Standard Deviation) in Points on the Likert Scale Between the 10th and the 90th Percentile of Experts	Mann–Whitney U-Test, p-value 9-Point Likert Scale (Recoded Data)§
Screening for anemia (n=14)	−0.02 (0.00)	0.13 (0.14)	−0.05 (0.03)	0.4–2.8	5.2 (0.8)	4.7 (1.2)	<0.001 (0.110)
Screening for parodontitis (n=16)	0.07 (0.08)	0.33 (0.31)	0.16 (0.16)	0.1–2.3	4.7 (1.5)	3.8 (1.2)	<0.001 (0.034)
Screening for lipid disorders (n=12)	0.15 (0.19)	0.35 (0.33)	0.24 (0.26)	0.1–1.6	4.2 (1.2)	3.7 (0.9)	0.201 (0.341)
Screening for and counseling on obesity (n=15)	0.26 (0.27)	0.34 (0.30)	0.30 (0.28)	0.0–2.1	3.5 (0.7)	3.7 (0.9)	0.197 (0.340)
Screening for and counseling on alcohol consumption (n=10)	0.34 (0.40)	0.22 (0.16)	0.28 (0.26)	0.0–1.5	4.2 (0.7)	4.4 (1.0)	0.264 (0.448)
Screening for age-related vision impairment (n=9)	0.29 (0.30)	0.39 (0.44)	0.34 (0.36)	0.0–1.8	4.8 (1.0)	4.0 (1.8)	0.047 (0.214)
Screening for glaucoma (n=8)	0.45 (0.49)	0.34 (0.32)	0.41 (0.41)	0.2–1.5	4.0 (0.4)	4.1 (1.4)	0.186 (0.139)

(*Continued*)

Table 1 (Continued).

Intervention (Number of Outcomes)	Agreement of Citizens 9-Point Likert Scale† (Recoded Data)§	Agreement of Experts 9-Point Likert Scale† (Recoded Data)§	Agreement Between Citizens and Experts 9-Point Likert Scale‡ (Recoded Data)§	Range of Mean Differences Between Citizens and Experts in Points on the 9-Point Likert Scale	Mean difference (Standard Deviation) in Points on the Likert Scale Between the 10th and the 90th Percentile of Citizens	Mean Difference (Standard Deviation) in Points on the Likert Scale Between the 10th and the 90th Percentile of Experts	Mann–Whitney U-Test, p-value 9-Point Likert Scale (Recoded Data)§
Screening for hearing impairment (n=7)	0.40 (0.46)	0.55 (0.60)	0.47 (0.51)	0.3–2.3	4.4 (0.7)	3.2 (1.5)	0.185 (0.551)
Screening for hepatitis C (n=9)	0.38 (0.38)	0.44 (0.48)	0.41 (0.43)	0.1–1.3	3.9 (0.8)	3.8 (1.0)	0.329 (0.398)
Screening for depression (n=12)	0.34 (0.38)	0.15 (0.11)	0.23 (0.22)	0.1–2.8	3.8 (0.9)	4.0 (0.9)	0.066 (0.698)
Counseling on physical activity (n=10)	−0.07 (−0.06)	0.02 (0.07)	0.03 (0.04)	0.0–1.2	4.1 (0.6)	3.2 (1.0)	0.515 (0.787)
Screening for vitamin D deficiency (n=15)	0.16 (0.17)	0.06 (0.02)	0.06 (0.04)	1.2–3.8	4.3 (0.9)	4.6 (0.9)	<0.001 (<0.001)
Screening for osteoporosis (n=8)	0.23 (0.25)	0.03 (0.01)	0.09 (0.08)	0.2–2.1	4.3 (1.0)	4.8 (0.8)	0.123 (0.356)
Screening for chronic kidney diseases and urinary tract infections (n=8)	0.39 (0.37)	0.17 (0.16)	0.24 (0.24)	0.3–3.6	3.5 (1.0)	4.4 (0.7)	0.020 (0.240)
Overall	0.24 (0.26)	0.36 (0.32)	0.28 (0.28)				<0.001 (0.003)

Notes: †Krippendorff's alpha; ‡Cube root product measure (CRPm); §Recoded data using 3 GRADE categories.

(p=0.09) and chronic kidney diseases and urinary tract infections (p=0.06). On average, the experts underestimated the outcomes' importance to the citizens by 11% (average mean citizens: 6.2; average mean experts: 5.3).

We observed exceptionally large differences in the mean ratings of the outcomes for screening for vitamin D deficiency (Table 1), with some values differing by more than 3 points (quality of life, prevention of immune system diseases, prevention of heart disease, increase in life expectancy, stroke prevention, prevention of diabetes, and cancer prevention) (range 1.2 to 3.8 points). Individual citizens' and experts' judgments about the outcomes' importance varied substantially across interventions. The mean differences in the points on the 9-point Likert scale between the 10th and the 90th percentile, which gives an indication of where 80% of the ratings fall, ranged from 3.5 to 5.2 among the citizens and from 3.2 to 4.8 among the experts (Table 1).

Another trend that can be observed (see the figures in the Supplementary material) is that both groups generally perceived the benefit-related outcomes as more important than those related to risks. This was more prominent among the citizens, who consistently rated all risk-related outcomes lower than those related to benefits. The experts rated the risk-related outcomes, such as overdiagnosis or overtreatment, higher than some of the benefit-related outcomes for the screening for vitamin D deficiency, osteoporosis, chronic kidney diseases, and urinary tract infection.

Expectedly, the second Delphi round of outcome ratings yielded results with higher agreements, as participants were influenced by the combined citizens' and experts' results that were reported back after the first Delphi round (Table 2). The overall agreement among the citizens rose to α=0.30 (recoded data α=0.30) and among the experts to α=0.53 (recoded data α=0.46; see Table 2). The intergroup agreement was CRPm=0.39 (recoded data CRPm=0.36), with an increase of 0.11 (0.09 recoded data) percentage points compared to the first round. The agreements between the citizens and experts were now above CRPm=0.5 for health outcomes of the following interventions: screening for glaucoma, hearing impairment, and hepatitis C. The mean differences between the citizens and experts in the points on the 9-point Likert scale tended to be lower than those in the first Delphi round, with maximum values ranging from 1.1–3.3. The mean differences in the points on the Likert scale between the 10th and the 90th percentile still ranged from 2.8 to 4.7 among

the citizens and from 2.1 to 3.8 among the experts, reflecting individual values and preferences. The citizens still rated the outcomes as more important than the experts across all interventions (p<0.01 when testing 9 Likert scores and 3 GRADE categories). The experts underestimated the outcomes' importance to citizens by 7% (average mean citizens: 5.7; average mean experts: 5.2).

Discussion

Our study assessed whether expert judgments about the importance of beneficial and harmful outcomes differ from citizen preferences when considering 14 interventions for the development of the PHE program. The agreements both within and between the citizen and expert groups tended to be poor to fair, with variations between the interventions and outcomes. Further exploring the data by comparing the mean differences between the citizens and experts for each outcome, we found that citizens, on average, attached a higher importance to all the outcomes than the experts (p<0.01 across all interventions), a finding also observed by Demyttenaere et al.[10] This confirms that experts underestimate the importance patients attach to outcomes, and that experts are not ideal proxies for citizens in outcome ratings.

Overall, it seems that citizens find it harder to discriminate outcomes on a scale and consider most outcomes very important. However, both groups show a substantial variability in the ratings, as indicated by the considerably large mean differences in the points on the Likert scale between the 10th and the 90th percentile. The observed low agreement among medical experts is an unexpected finding, as one would assume that a common expertise leads to higher agreement in outcome ratings.

Another observation is that although both the citizens and experts rated beneficial outcomes higher than harmful outcomes, the experts tended to show more awareness of harmful outcomes. For screening for depression, vitamin D deficiency, osteoporosis, and chronic kidney diseases or urinary tract infections, experts rated harmful outcomes higher than citizens, while citizens rated beneficial outcomes higher than experts. One explanation is that citizens could be less concerned with risks or less aware that risks can be a problem than experts within a screening and counseling intervention context. For example, the findings from a qualitative study on the information needs in colorectal cancer screening showed that some participants do not want to be informed about the risk of getting a false test result.[20]

Table 2 Agreements and Mean Differences Among and Between the Citizens and Experts in the 2nd Round

Intervention (Number of Outcomes)	Agreement of Citizens 9-Point Likert Scale† (Recoded Data)§	Agreement of Experts 9-Point Likert Scale† (Recoded Data)§	Agreement Between Citizens and Experts 9-point Likert Scale‡ (Recoded Data)§	Change in Agreement Between Citizens and Experts from 1st Round‡ (Recoded Data)§	Range of Mean Differences Between Citizens and Experts in Points on the 9-Point Likert Scale	Mean Difference (Standard Deviation) in Points on the Likert Scale Between the 10th and the 90th Percentile of Citizens	Mean Difference (Standard Deviation) in Points on the Likert Scale Between the 10th and the 90th Percentile of Experts	Mann–Whitney U-Test, p-value 9-point Likert Scale (Recoded Data)§
Screening for anemia (n=14)	0.11 (0.12)	0.33 (0.29)	0.17 (0.16)	0.22 (0.13)	0.2–2.8	4.3 (0.7)	3.2 (1.1)	0.003 (0.089)
Screening for parodontitis (n=16)	0.21 (0.21)	0.32 (0.31)	0.26 (0.26)	0.10 (0.10)	0.2–1.9	4.5 (1.0)	3.4 (1.2)	0.004 (0.101)
Screening for lipid disorders (n=12)	0.25 (0.17)	0.43 (0.34)	0.32 (0.25)	0.08 (−0.01)	0.0–1.5	4.2 (1.1)	3.1 (1.2)	0.272 (0.459)
Screening for and counseling on obesity (n=15)	0.14 (0.14)	0.56 (0.45)	0.27 (0.25)	−0.02 (−0.04)	0.1–2.2	4.7 (0.9)	2.9 (0.9)	0.331 (0.418)
Screening for and counseling on alcohol consumption (n=10)	0.18 (0.21)	0.49 (0.46)	0.29 (0.31)	0.01 (0.05)	0.0–1.4	4.7 (1.5)	2.9 (0.5)	0.410 (0.526)
Screening for age-related vision impairment (n=9)	0.42 (0.44)	0.61 (0.58)	0.48 (0.49)	0.14 (0.12)	0.2–3.3	3.4 (0.4)	2.7 (1.2)	0.092 (0.235)
Screening for glaucoma (n=8)	0.61 (0.58)	0.55 (0.53)	0.58 (0.56)	0.18 (0.15)	0.1–1.6	3.3 (0.7)	2.9 (0.9)	0.186 (0.159)

Screening for hearing impairment (n=7)	0.48 (0.48)	0.79 (0.73)	0.61 (0.58)	0.14 (0.07)	0.5–1.8	3.6 (0.4)	2.1 (1.0)	0.153 (0.551)
Screening for hepatitis C (n=9)	0.47 (0.46)	0.62 (0.61)	0.53 (0.53)	0.12 (0.10)	0.0–1.8	3.6 (0.6)	2.8 (1.1)	0.500 (0.535)
Screening for depression (n=12)	0.32 (0.37)	0.44 (0.34)	0.38 (0.35)	0.15 (0.14)	0.0–1.1	3.8 (0.7)	3.0 (0.7)	0.123 (0.767)
Counseling on physical activity (n=10)	−0.06 (−0.05)	0.07 (0.07)	−0.05 (−0.02)	−0.08 (−0.06)	0.2–1.1	3.1 (0.5)	3.4 (0.9)	0.546 (0.766)
Screening for vitamin D deficiency (n=15)	0.33 (0.31)	0.09 (0.07)	0.16 (0.13)	0.10 (0.09)	0.4–3.1	3.6 (0.8)	3.8 (0.7)	<0.001 (0.002)
Screening for osteoporosis (n=8)	0.51 (0.49)	0.24 (0.23)	0.36 (0.35)	0.27 (0.27)	0.1–1.7	3.1 (1.1)	3.3 (1.2)	0.185 (0.247)
Screening for chronic kidney diseases and urinary tract infections (n=8)	0.57 (0.57)	0.43 (0.31)	0.49 (0.41)	0.24 (0.18)	0.1–2.9	2.8 (1.2)	2.7 (0.9)	0.057 (0.118)
Overall	0.30 (0.30)	0.53 (0.46)	0.39 (0.36)	0.11 (0.09)				<0.001 (0.012)

Notes: [†]Krippendorff's alpha; [‡]Cube root product measure (CRPm); [§]Recoded data using 3 GRADE categories.

We observed the biggest difference in the mean outcome ratings between the citizens and experts for screening for vitamin D deficiency, with the mean differences across outcomes ranging from 1.2–3.8 (Table 1 and Supplementary material Figure S12). This indicates that the citizens perceived vitamin D deficiency as a much more important health threat than the experts, which could be a result of the knowledge difference between the two groups. While experts rated prevention of osteoporosis and fractures much higher than other beneficial outcomes in relation to vitamin D deficiency, citizens considered a range of beneficial outcomes, such as stroke and cancer prevention, as important. Similar ratings were observed for screening for osteoporosis, where experts consider the prevention of bone fractures as most important, and citizens rate all the beneficial outcomes as similarly important.

We further noticed that the agreement within and between groups can be low, even though the mean values of the outcome ratings indicate little difference between citizens and experts, for example, for counseling on physical activity (Table 1 and Supplementary material Figure S11). In contrast, interventions with a moderate or fair agreement, such as screening for glaucoma and for hearing impairment (Table 1), show notable differences in the mean values between citizens and experts for some outcomes (eg, improvement in well-being, reduction in eye pain, memory improvement, prevention of dizzy spells; Supplementary material Figures S7 and S8). This indicates that mean values are of little informative value if the agreement is low, but if the agreement is high, it is still important to look at potential differences in individual outcomes.

The findings support the application of a Delphi approach using at least two rating rounds during guideline development.[16] The second rating round increased the agreement levels, with four interventions' agreement levels advancing to a good strength of agreement (Table 2). Online approaches to engage participation in a guideline development process are convenient and facilitate greater openness and honesty among patients as opposed to face-to-face meetings, where patients may feel intimidated by clinicians and researchers. Potential disadvantages include the difficulty of engaging specific patient populations.[21] Khodyakov et al have developed a practical guidance for conducting online modified Delphi panels, which was published after this study was conducted.[22]

Our additional analysis of recoding the data into three instead of nine categories on the scale, which is eventually applied in the GRADE decision-making process,[8] showed that the agreement levels only slightly improved or even diminished when using three categories. This confirms that Krippendorff's alpha and the CRPm are robust measures toward recategorization, and that the current GRADE approach of using nine categories should not be changed.

On a larger scale, the findings of this study enhance our understanding of the impact of patient engagement on guideline development. They show that citizens attach a different importance to intervention outcomes in the context of PHE than experts, thereby influencing the decision of which interventions to include in the PHE program. This study has also advanced the research evidence on outcome selection and prioritization, which is integrated in Step 4 (Framing the question) of the framework proposed by Armstrong et al.[7] Their findings also emphasize the importance of citizen or patient participation using formal methods of involvement such as surveys or Delphi approaches.[16] Several researchers have criticized the roles of patients as often being symbolic rather inclusive.[23,24]

The strengths of this study include the use of agreement measures for comparison. Other studies have merely relied on the measure of means to compare the data.[9,10,25] Another strength is the use of plain language in the surveys sent to both the citizens and experts. We collaborated with science journalists who translated medical terms into plain language.

A limitation of this study was the small sample size (11 to 14 citizens and 9 to 10 experts) and the low response rate among citizens (37–47%). Estimating sample sizes for agreement studies can be a complicated undertaking,[26] with no simple guidance for the sample size requirements of the CRPm yet available. The strength of agreement is influenced by various factors, including the number of subjects or categories or the distribution of subjects among categories. Consequently, an agreement value of 0.5 based on 100 participants provides a much stronger message about the extent of the agreement among raters than an agreement value of 0.5 based on 10 participants.[27] There are also no guidelines or recommendations on the appropriate sample size for Delphi studies, and previous studies have been conducted with virtually any sample size, ranging from 10 to 100 or more participants.[28] We wanted to get some idea of the potential differences in the outcome ratings between citizens and experts in a screening or counseling intervention context that occur if a relatively small number of stakeholders are involved. It was, therefore, beyond the scope of this work to achieve a sample size with adequate statistics for an agreement analysis. However, we do recognize the exploratory nature and limited generalizability of our study and propose that future research considers a larger sample size.

Conclusions

This study adds empirical knowledge to our current methodological understanding of guideline development processes in a preventive context. The differences and variability of the results between the citizens and experts demonstrate the importance of citizens' or patients' involvement in outcome selection and the prioritization of a guideline development process using formal methods such as surveys or Delphi approaches. These differences show that experts are not ideal proxies for citizens in outcome ratings.[23,24] The large variability among the ratings within each group further suggests involving larger groups of patients or citizens and also experts in outcome ratings. However, further research including a statistically powerful sample size must be undertaken to confirm this conclusion.

Abbreviations

AGREE, Appraisal of Guidelines for Research and Evaluation; CRPm, cube root product measure; GRADE, Grading of Recommendations Assessment, Development and Evaluation; PHE, periodic health examination.

Data Sharing Statement

The survey developed for this study was conducted in the German language. A copy can be requested from the corresponding author. The datasets analyzed during the current study are not publicly available due to privacy or ethical restrictions but are also available from the corresponding author on reasonable request.

Ethics Approval and Informed Consent

Ethical approval was granted by the Danube University Research Ethics committee (EK GZ 16/2015–2018). We sought informed consent in the form of signatures from the participants of the focus groups and implied consent from the participants of the survey. The Danube University Research Ethics committee approved both forms of consent.

Acknowledgments

Author Contributions

Funding

Disclosure

The authors declare that they have no conflicts of interest in this work.

References

1. Roman BR, Feingold J. Patient-centered guideline development: best practices can improve the quality and impact of guidelines. *Otolaryngol Head Neck Surg*. 2014;151(4):530–532. doi:10.1177/0194599814544878

2. van Dulmen SA, Lukersmith S, Muxlow J, Santa Mina E, Nijhuis-van der Sanden MW, van der Wees PJ. Supporting a person-centred approach in clinical guidelines. A position paper of the Allied Health Community - Guidelines International Network (G-I-N). *Health Expect*. 2015;18(5):1543–1558. doi:10.1111/hex.12144

3. Institute of Medicine (US) Committee on Standards for Developing Trustworthy Clinical Practice Guidelines. *Clinical Practice Guidelines We Can Trust*. Washington (DC): National Academies Press (US); 2011.

4. Guideline International Network. G-I-N PUBLIC toolkit: patient and public involvement in guidelines. Published 2015. Available from: http://www.g-i-n.net/working-groups/gin-public/toolkit. Accessed July 18, 2019.

5. Brouwers MC, Kerkvliet K, Spithoff K. The AGREE reporting checklist: a tool to improve reporting of clinical practice guidelines. *BMJ*. 2016;352:i1152. doi:10.1136/bmj.i1152

6. Selva A, Sanabria AJ, Pequeño S, et al. Incorporating patients' views in guideline development: a systematic review of guidance documents. *J Clin Epidemiol*. 2017;88:102–112. doi:10.1016/j.jclinepi.2017.05.018

7. Armstrong MJ, Rueda JD, Gronseth GS, Mullins CD. Framework for enhancing clinical practice guidelines through continuous patient engagement. *Health Expect*. 2017;20(1):3–10. doi:10.1111/hex.12467

8. Schünemann H, Brożek J, Guyatt G, Oxman A Introduction to GRADE Handbook. Handbook for grading the quality of evidence and the strength of recommendations using the GRADE approach. GRADE Working Group; Published October, 2013. Available from: http://gdt.guidelinedevelopment.org/app/handbook/handbook.html. Accessed March 21, 2017.

9. Gutman T, Lopez-Vargas P, Manera KE, et al. Identifying and integrating patient and caregiver perspectives in clinical practice guidelines for percutaneous renal biopsy. *Nephrology*. 2019;24(4):395–404. doi:10.1111/nep.13406

10. Demyttenaere K, Donneau AF, Albert A, Ansseau M, Constant E, van Heeringen K. What is important in being cured from depression? Discordance between physicians and patients (1). *J Affect Disord*. 2015;174:390–396. doi:10.1016/j.jad.2014.12.004

11. Janssen IM, Scheibler F, Gerhardus A. Importance of hemodialysis-related outcomes: comparison of ratings by a self-help group, clinicians, and health technology assessment authors with those by a large reference group of patients. *Patient Prefer Adherence*. 2016;10:2491–2500. doi:10.2147/PPA.S122319

12. Alonso-Coello P, Montori VM, Díaz MG, et al. Values and preferences for oral antithrombotic therapy in patients with atrial fibrillation: physician and patient perspectives. *Health Expect*. 2015;18 (6):2318–2327. doi:10.1111/hex.12201

13. Gesundheit.gv.at. Die Vorsorgeuntersuchung auf einen Blick. Published 2019. Available from: https://www.gesundheit.gv.at/leben/gesundheitsvorsorge/vorsorgeuntersuchung/was-wird-gemacht. Accessed July 18, 2019.

14. World Medical Association. World medical association declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA*. 2013;310(20):2191–2194. doi:10.1001/jama.2013.281053

15. Sommer I, Titscher V, Gartlehner G. Participants' expectations and experiences with periodic health examinations in Austria - a qualitative study. *BMC Health Serv Res*. 2018;18(1):823. doi:10.1186/s12913-018-3640-6

16. Helmer-Hirschberg O. *Analysis of the Future: The Delphi Method*. California, USA: RAND Corporation; 1967.

17. Panda M, Paranjpe S, Gore A Measuring intergroup agreement and disagreement. Cytel Statistical Software & Services Private Limited; Published 2018. Available from: https://arxiv.org/ftp/arxiv/papers/1806/1806.05821.pdf. Accessed March 29, 2019.

18. Altman DG. *Practical Statistics for Medical Research*. London: Chapman and Hall; 1991.

19. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Published 2018. Available from: http://www.R-project.org. Accessed March 29, 2019.

20. Kirkegaard P, Mortensen GL, Mortensen SL, Larsen MB, Gabel P, Andersen B. Making decisions about colorectal cancer screening. A qualitative study among citizens with lower educational attainment. *Eur J Public Health*. 2016;26(1):176–181. doi:10.1093/eurpub/ckv207

21. Grant S, Hazlewood GS, Peay HL, et al. Practical considerations for using online methods to engage patients in guideline development. *Patient*. 2018;11(2):155–166. doi:10.1007/s40271-017-0280-6

22. Khodyakov D, Grant S, Denger B, et al. Practical considerations in using online modified-Delphi approaches to engage patients and other stakeholders in clinical practice guideline development. *Patient*. 2020;13(1):11–21. doi:10.1007/s40271-019-00389-4

23. van de Bovenkamp HM, Zuiderent-Jerak T. An empirical study of patient participation in guideline development: exploring the potential for articulating patient knowledge in evidence-based epistemic settings. *Health Expect*. 2015;18(5):942–955. doi:10.1111/hex.12067

24. Duffett L. Patient engagement: what partnering with patient in research is all about. *Thromb Res*. 2017;150:113–120. doi:10.1016/j.thromres.2016.10.029

25. Janssen IM, Gerhardus A, Schroer-Gunther MA, Scheibler F. A descriptive review on methods to prioritize outcomes in a health care context. *Health Expect*. 2015;18(6):1873–1893. doi:10.1111/hex.12256

26. Krippendorff K. Agreement and information in the reliability of coding. *Commun Methods Meas*. 2011;5(2):93–112. doi:10.1080/19312458.2011.568376

27. Gwet K. *Handbook of Inter-Rater Reliability*. Maryland, USA: Advanced Analytics Press; 2014.

28. Akins RB, Tolson H, Cole BR. Stability of response characteristics of a Delphi panel: application of bootstrap data expansion. *BMC Med Res Methodol*. 2005;5:37. doi:10.1186/1471-2288-5-37