Structured Expert Judgment Findings Informing World Health Organization Foodborne Disease Estimates 2026

Gabriela F. Nane[1], Tine Hald[2], Willy Aspinall[3], Roger M. Cooke[1,4], Arie H. Havelaar[5], Yuki Minato[6], Charlee Roberts[6], Lapo Miughini-Gras[7,8], Sandra Hoffmann[9], Kunihiro Kubota[10], Shannon E. Majowicz[11], Martyn D. Kirk[12], Teresa Estrada Garcia[13], Lucy J. Robertson[14], Paul R. Torgerson[15], Karen H. Keddy[16], Lea S. Jakobsen[2], Antonio Agudo[17], Bodille Blomaard[1], Alessandra Primavera[1], Brecht Devleesschauwer[18,19], Sara M. Pires[2]

Affiliations

1. Delft Institute of Applied Mathematics, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Mekelweg 4, 2628 CD, Delft, the Netherlands
2. National Food Institute, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark
3. Aspinall Consulting Ltd., Tisbury, Wiltshire SP3 6HF, United Kingdom; School of Earth Sciences, Bristol University, Bristol BS8 1RJ, United Kingdom
4. Resources for the Future (ret), Washington DC, Untied States of America
5. Emerging Pathogens Institute, Global Food Systems Institute, Department of Animal Sciences, University of Florida, 2055 Mowry Rd, Gainesville, Florida, 32610, United States of America
6. Department of Nutrition and Food Safety, World Health Organization, 20 Avenue Appia, 1211 Geneva, Switzerland
7. National Institute for Public Health and the Environment (RIVM), Center for Infectious Disease Control, Antonie van Leeuwenhoeklaan 9, 3721 MA, Bilthoven, the Netherlands
8. Utrecht University, Faculty of Veterinary Medicine, Yalelaan 2, 3584 CL, Utrecht, the Netherlands.
9. George Washington University, School of Public Health, Washington, DC, United States of America
10. National Institute of Health Sciences, Division of Food Safety Information, 3-25-26 Tonomachi, Kawasaki-ku, Kawasaki-City, Kanagawa, Japan
11. School of Public Health Sciences, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada, N2L 3G1
12. National Centre for Epidemiology and Population Health, Australian National University, Acton, Australian Capital Territory 2601, Australia
13. Department of Molecular Biomedicine, CINVESTAV-IPN, Mexico City, Mexico
14. Parasitology, Department of Paraclinical Sciences, Faculty of Veterinary Medicine, Norwegian University of Life Sciences, Ås, Norway
15. Section of Epidemiology, Vetsuisse Faculty, University of Zürich, Zürich, Switzerland
16. Department of Veterinary Tropical Diseases, University of Pretoria
17. Unit of Nutrition and Cancer, Cancer Epidemiology Research Program, Catalan Institute of Oncology-ICO, and Nutrition and Cancer Group, Bellvitge Biomedical Research Institute-IDIBELL. L'Hospitalet de Llobregat, Barcelona, Spain
18. Department of Epidemiology and Public Health, Sciensano, Rue J Wytsman 14, 1050 Brussels Belgium
19. Department of Translational Physiology, Infectiology and Public Health, Ghent University, Salisburylaan 133, 9820 Merelbeke, Belgium

1. Introduction

Robust estimates of the global burden of foodborne diseases are essential for guiding food safety policy and prioritizing interventions. The release of the first World Health Organization (WHO) global burden estimates in 2015 (WHO, 2015; Havelaar et al., 2015) represented a major step toward quantifying the public health impact of contaminated food and highlighted substantial gaps in both surveillance data and knowledge about how different hazards are transmitted. To continue strengthening the evidence base for food safety decision-making, WHO reconvened the Foodborne Disease Burden Epidemiology Reference Group (FERG) for 2021-2025 with a mandate to update the global estimates, expand analyses to the national level, and develop indicators for monitoring progress in reducing foodborne disease (WHO, 2022).

A central component in estimating the burden of foodborne diseases is the attribution of illnesses to major transmission pathways, including food, water, the environment, and contact with humans or animals, and, within the food pathway, to specific food categories (Pires et al., 2009; Mughini-Gras et al., 2019). Attribution remains challenging because many hazards that contribute to the global burden of foodborne disease are not exclusively foodborne, and the relative importance of transmission routes varies widely across hazards, regions, seasons, and food systems (Hald et al., 2016; Hoffmann et al., 2017). Although several methodological approaches have been developed for source attribution, ranging from analyses based on outbreak investigations and surveillance data to intervention studies, microbial subtyping, exposure assessments, and probabilistic modelling (Pires et al, 2009; EFSA BIOHAZ Panel, 2013), empirical data suitable for comprehensive source attribution analyses remain limited. This is particularly true on the global scale and for hazards with sparse surveillance data, heterogeneous epidemiology, or limited monitoring information.

While data-driven approaches are preferred when high-quality datasets exist, they remain applicable only to a subset of hazards and countries (Crotta et al., 2022; Davydova et al., 2025). For many globally important hazards, the absence of consistent, comparable, or representative empirical data precludes the use of such methods and necessitates alternative approaches.

Structured expert judgment (SEJ) offers a systematic and transparent way to quantify uncertainty and synthesize the available evidence where empirical data are insufficient (Aspinall & Cooke, 2013). SEJ has been used in multiple domains of risk assessment (Baxter et al., 2008; Bamber et al., 2013; Hanea et al., 2021), including within food safety to estimate attribution fractions and parameterize models under conditions of high uncertainty (e.g., Butler et al., 2015; Beshearse et al., 2021; Sapp et al., 2022). Within FERG for 2007-2015, the Source Attribution Task Force (SATF) advised the SEJ in 2015 to produce the first global estimates of the proportion of illnesses attributable to food and specific food categories (Hald et. al, 2016, Hoffmann et al., 2017, Aspinall et al., 2016). Building on that earlier work, WHO decided to commission a new study, and after the open call[1], the team at the Delft University of Technology was commissioned to carry out a new SEJ study under the advice of the FERG and especially SATF, to inform the updated WHO global burden of foodborne diseases 2026 edition.

This paper provides a detailed technical description of the methods and findings of the 2026 SEJ study commissioned by WHO. We present the design and implementation of the elicitation, including the construction of the calibration and target questions, expert identification, selection and training, performance-based weighting, and statistical aggregation procedures. We also report updated global attribution estimates for major transmission pathways and food categories for 40 enteric, parasitic, and chemical hazards. These findings formed a key input into the updating of the WHO foodborne disease burden estimation framework and represent the most comprehensive evidence to date on the global distribution of exposure routes for foodborne diseases. For more information on how these estimates

[1] https://cdn.who.int/media/docs/default-source/foodborne-diseases/ferg/ferg-satf-001-tor.pdf?sfvrsn=3ade1f0f_3

were used to inform the WHO attribution of burden of foodborne disease and a discussion of the findings in the context of other studies and empirical data, we refer to Pires et. al (2026).

2. Structured expert judgment

This structured expert judgment study considered 40 hazards: 24 enteric hazards (including 14 diarrhoeal diseases hazards and 10 invasive and intoxicating hazards), 10 parasites and 6 chemical hazards. The list of the hazards, as well as the acronyms used for the analysis are included in Table A1 in Appendix A.

The six standard WHO regions (African Region, Region of the Americas, South-East Asia Region, European Region, Eastern Mediterranean Region, and Western Pacific Region) were divided into 17 clusters of countries (subregional clusters or subregions). This constituted a revision of the WHO FERG 2015 subregions, which was required in the light of subsequent changes in mortality patterns and other, newer epidemiological insights, as well as to account for economic aspects, as captured by the revised World Bank income classifications. The list of countries assigned to each of the subregions is included in Table A2 (Appendix A) and a map of the 17 subregions can be found in Appendix A (Figure A1). The correspondence between the subregions in this study and those used for the WHO 2015 estimates (Hald et al., 2015) can be found in Figure A2 in Appendix A.

The following sub-sections will present details regarding the design of elicitation questions, expert identification and selection, their training, and how the elicitation was carried out.

2.1. Target questions

Experts were asked to estimate the proportion of infections attributed to six major transmission pathways: food, contact with animals, human contact, water, soil and 'other'; for *Trypanosoma cruzi* vector-borne transmission was considered an additional pathway. The pathway definitions are included in Table A4 in Appendix A. Within the proportion foodborne, the experts were asked to attribute disease to 14 specific food groups: beef, small ruminants' meat, pigs' meat, poultry meat, game, dairy, eggs, vegetables, fruits and nuts, grains and beans, oils and sugar, finfish, shellfish (including snails) and seaweed. The experts were asked to consider the point of attribution as the point where the food entered the place of preparation just before consumption (e.g., 'kitchen door'). Uncertainty about the source attribution estimates was elicited from experts in the form of point estimates (50^{th} percentile of their subjective distribution, or the median), together with a 90% confidence interval, specified by their 5^{th} percentile and their 95^{th} percentile values of their subjective distribution. Experts self-selected the hazards and sub-regions for which they wished to provide these probabilistic assessments.

For certain hazards, specific pathways or food groups were considered (biologically) implausible and therefore appeared as 'blocked' in the questionnaire presented to the experts. The list of the blocked pathways/food groups is reported in Tables A5.1.- A5.5 in Appendix A. To ensure a consistent framework for all questions, and to allow all datasets to be uniformly processed by the software program, the three-point format (5^{th}, 50^{th} and 95^{th} percentile) was retained for blocked pathways/food groups, these being ascribed the pre-specified values 10^{-8}, 10^{-6}, 10^{-4}, to reflect the fact that such infection is virtually impossible. Experts were informed of the blocked pathways/food groups during their elicitations. If they considered that any of the pathways/food groups should not be regarded as blocked, they could change these pre-ascribed values.

Together with the quantitative source attribution judgments, the experts were asked to provide additional qualitative input. Firstly, they were asked to specify what they had in mind for the pathway 'other' when assessing this unspecified pathway category. Possible alternatives were presented when

the category was listed in the elicitation protocol, such as 'airborne', 'pollution', or 'occupational'. Finally, for each specific hazard/subregion, the experts were asked to reflect on and record the main factors that they considered when providing their judgment(s).

For nineteen of the 40 hazards, source attribution estimates had been obtained using the same expert elicitation methodology, which are known as the 2015 attribution estimates (see Hald et al., 2016). However, given the updates in the number and distribution of subregions, as well as changes in the compositions of many of the subregion groupings, those earlier estimates could not be compared exactly with the newer 2026 attribution estimates. Nonetheless, the 2015 attribution estimates corresponding to the new subregions were mapped, based on population sizes in 2015, as reported by the World Bank. The subregional mapping is specified in Figure A2 in Appendix A.

To ensure that the experts in the present study were informed about the 2015 findings, and to probe their views on those previous results, the 2015 attribution estimates, re-calculated to match the new subregions were presented to the experts as plots of major pathway estimates (median and 95% confidence interval). The experts were asked also whether their view would have been substantially different, had they been on the former panel. More specifically, they were asked if there have been any changes from 2010 (reference year for the 2015 estimates) to 2019 (reference year for the 2026 estimates) to pertinent influences or factors that might have occurred which could have substantially altered the 2015 hazard estimates in the intervening period. If they thought so, they were invited to reflect on the nature of such potential influences or factors. To support the experts' reasoning, the SEJ study team gathered a list of general considerations for potential factors (Table A3 in Appendix A). The experts were encouraged to consider other potential factors, if relevant.

For the 21 hazards not included in the 2015 attribution estimates, experts were invited to reflect on whether there have been changes that could have substantially altered source attribution estimates of these hazards from 2010 to 2019. These qualitative insights were all marked as optional when requested on the experts.

2.2. Calibration questions

We employed the Classical Model for SEJ, which employs calibration questions (or variables) to calibrate the mathematical model using the aggregate experts' uncertainty assessments (Cooke, 1991). Calibration questions are questions for which the realization is known (or will be known within the study period) to the analyst team, but not to the experts. When realizations are not known yet during the elicitation, the calibration questions are referred to as predictions. When the realizations are known at the moment of the elicitation, then the calibration questions are referred to as retrodictions. Though not ideal, when predictions are not accessible, relevant subject matter retrodiction remain the best available solution (Hanea & Nane, 2021). In this study, we used retrodictions only.

Calibration questions covered five main topics relevant for source attribution: 1) food supply (all panels), 2) health and diarrhoeal disease (enteric and parasitic panels), 3) disease surveillance (enteric and parasitic panels), and 4) WASH – water and sanitation (enteric and parasitic panels) and 5) exposure and epidemiology of chemicals (chemical panels). The questions were developed using data from public statistics, scientific reports, and peer-reviewed literature. Within each topic, questions were tailored both to the broad class of hazards, i.e., enteric, parasitic, or chemical, and to geographical regions. Food supply questions were tailored for each region.

Topics related to health and diarrhoeal disease, disease surveillance, and WASH were included only for enteric and parasitic hazards. Although *Cryptosporidium* spp., *Giardia* spp., *Entamoeba histolytica* and *Cyclospora* were classified in the study as enteric hazards, the experts eliciting these pathogens received the same calibration questions as for the other parasites. Disease surveillance questions covered only a

subset of hazards, as tailoring questions for each of the 34 hazards was not feasible. Moreover, disease surveillance questions were further tailored to the regional context. The WASH-related questions were identical across enteric and parasitic hazards.

For chemical hazards, further tailoring was designed specifically for aflatoxins B1, Dioxin & DL-PCBs and heavy metals.

In total, 41 calibration questions were developed by the study team, and every hazard/sub-region combination yielded 13 calibration questions. The list of all questions' classification in terms of the abovementioned topics and regions is included in Tables B1.1.-B1.3. in Appendix B and the list of all 41 questions is presented in Table B2 in Appendix B. During the elicitation, experts' choice of hazards/sub-regions triggered the automatic selection of calibration questions. Selection of multiple hazards/sub-regions entailed the cumulation of calibration questions, so that the complete set of calibration questions was ensured for each hazard/sub-regions combination. Consequently, if the experts chose to assess enteric and/or parasitic hazards within different regions, they needed to answer food supply questions for each separate question.

2.3. Expert training

Comprehensive training materials were prepared and circulated to all expert participants. Four videos of about 5 minutes each were recorded to introduce: (1) the study; the (2) the Classical Model (3) providing uncertainty estimates, and (4) a role play reflecting typical queries and concerns that experts have during an elicitation. The links to the videos are included in Table B3 in Appendix B. Experts could also follow a training exercise answering questions to enhance familiarisation. This consisted of three generic questions regarding food accessibility, and three more specific questions, depending on their broad expertise (e.g., enteric/parasitic or chemicals hazards). The list of the training questions can be found in Table B2 in Appendix B.

After each question, the experts were provided with feedback regarding the known realization values and whether their uncertainty assessments captured the realizations. After answering all the training questions, the experts were provided with feedback for all questions and asked to reflect on how many of their 90% uncertainty intervals captured the realizations. In case the realizations were not captured at the expected rate for 'good calibration', they were invited to consider increasing their uncertainty intervals when providing assessments. Finally, five practice questions regarding the Classical Model were asked. These concerned characteristics of the uncertainty assessments, such as the percentiles needing to be in a strictly increasing order, or that the 5^{th} and the 95^{th} percentiles needed not to be symmetric around the 50^{th} percentile. The training module was expected to take maximum 45 minutes to complete.

The training module was implemented in Qualtrics survey software (Qualtrics, Provo, UT). As required by the problem owner (WHO), experts were asked to complete and sign a consent form, which described the purpose of the study, the procedure and duration, what data were going to be used and how, the risks and benefits of participating in the study, as well as confidentiality aspects and the rights of the experts. By signing the form, the experts declared their voluntary consent to participate in the study. Experts were also required to provide a declaration covering potential conflicts of interest regarding any financial, professional or other interests or activity relevant to this study.

2.4. Expert identification and selection

On 10 July 2023, the WHO issued a public call inviting experts in food safety and related disciplines, including microbiology, water quality, environmental health, and hygiene, to express their interest in

participating in this SEJ study. In addition, further candidates were identified through snowball sampling and outreach within the professional networks of WHO and members of the new FERG.

The evaluation and selection of experts were carried out jointly with the WHO Collaborating Centre for Risk Assessment in Food and Water, hosted at the National Institute for Public Health and the Environment (RIVM) in the Netherlands. Candidates were assessed based on their curriculum vitae, publication records, and professional background, according to criteria established by the SATF. Selection was guided by documented subject-matter expertise relevant to the domains under consideration. Separate evaluation criteria were applied to academic experts (e.g., university faculty and researchers) and to non-academic professionals (e.g., government officials, food safety inspectors, and other practitioners), as outlined in Table C1 in Appendix C. Experts employed in the commercial or industrial sector were not included in the selection process.

The screening and selection of experts was done in collaboration with the WHO Collaborating Center for Risk Assessment in Food and Water, hosted by the National Institute for Public Health and the Environment of the Netherlands (RIVM), based on a review of submitted curriculum vitae (CV) and publication lists using criteria defined by the SATF (Pires et al., 2026). The selected criteria included level and relevance of university degree; years of experience; and relevant publications on the hazards and regions they self-selected. Experts from academic or governmental institutions were evaluated separately. The selection aimed to involve experts based on their domain-specific knowledge as documented in their CV.

After experts were selected, the study team informed the experts whether they could contribute to the study.

In addition to the general information the experts had provided already, they were invited to answer a Qualtrics survey to provide more specific information regarding their hazard/subregional expertise. Experts self-reported their level of expertise, with values from 1 (low; no direct experience), 3 (medium; some direct experience, to 5 (high; the hazard is the primary focus of their professional work. Using this grading guidance, experts could provide the level of their self-reported expertise for each hazard/subregion combination and indicate whether their level of expertise varied across subregions.

The qualitative self-assessments allowed the study team to identify hazards for which experts reported higher than medium expertise, and to detect expertise gaps across hazards and subregions. Active scouting was conducted by WHO to address these gaps.

2.5. Elicitor selection and training

In an expert judgment study, an elicitor or facilitator ensures a smooth elicitation process and provides support (not coaching) during an elicitation. Knowledge about the study, as well as about the method is important. Within this study, we ensured that the elicitors had sufficient knowledge about the study and were provided with extensive training and support on structured expert judgment.

WHO made an initial call for elicitors to support the study. Prospective candidates were queried regarding education and professional experience, experience with communication in a professional environment, as well as access to a stable internet connection and a quiet environment. To ensure support for experts who might feel more comfortable with non-English languages, prospect elicitors were asked if they were able to conduct the elicitations in French, Spanish, Arabic, Mandarin, Portuguese or Russian. A total of 64 elicitors were selected, representing full regional and time zone coverage. A list with the elicitors who conducted at least one elicitation is included in Table C1 in Appendix C.

Elicitors received the same training materials as the experts to familiarize themselves with the study objectives, the target and calibration questions, and the expert judgment method. Two online training sessions were held specifically for the elicitors, where additional information was provided and elicitors could ask questions. Preparatory documents were provided to the elicitors, including the list of calibration questions, information about the hazards, the transmission pathways and the food groups.

A stock script was prepared to support elicitors when they conducted individual interviews and were prompting experts for their quantitative judgments. The main purpose of this script was to ensure as much uniformity as possible across the whole elicitation – given its many combinations of hazards and subregions, and consequently elicitation panels.

2.7. Elicitation

One-to-one online interviews with the experts were organised and conducted by the trained elicitors. The elicitors were put in contact with experts who were in the same region or similar time zone. During the interview, the experts answered the calibration questions. If time allowed, the experts also assessed one hazard/subregion. Otherwise, the elicitor introduced the target questions to familiarize the expert with the type of questions. The experts were asked to provide the assessments for all target questions within two weeks from the online session. The experts' follow-up was done by the study team.

At the beginning of the interview, the experts were reminded of the study purpose and the structured expert judgment method. The Qualtrics link that contained the calibration questions was shared with the experts. Within the Qualtrics survey, experts provided again their hazard/subregional expertise. The elicitors ensured that the expert still met the required self-reported expertise level (4 or 5) for the hazard/subregion combination selected.

The choices of specific hazards/subregions led to automatically tailored calibration questions, as described in Section 2.2. Thirteen calibration questions were given for each broad class of hazards and subregions. If experts selected multiple subregions across different regions, they needed to respond to more questions for food supply and outbreak and disease surveillance (see Table B2 in Appendix B).

For any hazard, the given pathways were assumed exhaustive and mutually exclusive, meaning that the estimated means of the obtained distributions should ideally sum to 100%. This condition is not expected to be satisfied by the estimated medians; that is, the estimated medians may not necessarily add to 100%. Nonetheless, since they represent the relative contribution of all major pathways, it is expected that they will add to something close to 100%. Experts were reminded that there is no theoretical statistical requirement for their estimated 5^{th} and 95^{th} percentiles of exhaustive and mutually exclusive categories to sum to totals near 100%.

The Qualtrics survey was coded as to indicate when an exceptionally low or high median sum value was determined from the pathways' or food groups' assessed medians, via a message that indicated the sum was quite far from 100%. After receiving this message, the expert was invited to reconsider and perhaps update their set of assessments as needed.

2.8. Mathematical validation and aggregation

We employed the Classical Model for Structured Expert Judgment (Cooke, 1991) to validate and aggregate experts' uncertainty assessments. A detailed description of the method is included in Appendix D. For more details, see, e.g., Hanea and Nane (2021).

The model proposes a validation framework to aggregate experts' assessments. The framework relies on the set of calibration questions and two objective measures of performance to derive aggregation weighting: *statistical accuracy* (also referred to as the calibration score) and *information*. Loosely, statistical accuracy measures the statistical likelihood that a set of experimental results correspond, in a statistical sense, with an expert's assessments. More precisely, under the Classical Model, statistical accuracy is scored as the p-value at which we would falsely reject the hypothesis that expert's probability statements were statistically accurate. Loosely, the information in a distribution is the degree to which the distribution is concentrated. An information score is computed for each variable (question), and the average information score reflects the overall informativeness of expert's assessments. A combined score is obtained by multiplying the statistical accuracy and informativeness to reflect the overall performance of quantifying uncertainty.

Combined scores are used to derive weighting schemes to combine experts' uncertainty distributions for the target questions. The resulting distributions are referred to as Decision Makers (DMs). Global weights, when the overall information score is employed for each expert lead to PW Decision Maker. In contrast, item weights are question specific and account for experts' different informativeness across questions. They lead to an Item Weight (IW) Decision Maker. An equal weight (EW) Decision Maker can also be considered. The different DMs can also be applied to the calibration questions and evaluated with respect to the statistical accuracy and informativeness. Moreover, an optimization procedure can be employed to yield the best performing DM, with respect to the combined score, by introducing a threshold indicator for experts' statistical accuracy. The resulting DMs are referred to as PW_opt, when global weights are used, and IW_opt, when item weights are used.

As recommended by the EU/USNRC Procedures Guide, we considered a minimum of 4 expert assessments per hazard/subregion for subregional tailored estimates (Cooke and Goossens, 2000).

2.9. Regional, economic and global proxies

Despite the huge effort to ensure such regional expertise coverage for all hazards and subregions, the goal of having at least four distinct expert assessments for all hazards/subregions combinations was extremely challenging. Consequently, three types of proxies were defined to handle cases where this was not achieved.

Regional proxies

First regional proxies were considered for imputation at the regional level. The necessary condition for regional proxies was to have at least four assessments for the entire broad WHO region. Depending on the expert data availability in the rest of the subregions, experts' data were merged. In general, the A subregions (especially AMR A, EUR A and WPR A) were much better represented than other subregions[2]. Similarly, B or BC subregions were better represented than C or D subregions. Consequently, the merging resulted in proxies leaning towards a higher income subregion (from B to A, from C to B, etc.). An important exception from this 'rule' was done when subregion A had considerably more assessments than region B and C combined. For example, subregion A had assessments from 22 assessments, while subregion B and C had three and two assessments, respectively. In such a scenario, it was decided to merge assessments from subregions B and C, as merging these subregional assessments with those from subregion A would unduly influence the much smaller panels estimates.

[2] An exception was AFR AB and EMR A, which have been less represented than AFR C and EMR BC.

Examples of regional proxies include merging assessments from EUR C and EUR B, and from AFR AB with AFR C. A full description of the regional merging procedure is included in Table E1 in Appendix E.

Economic proxies

When the application of regional proxies was not sufficient for aggregation, i.e., less than four expert assessments were available for an entire region, then economic proxies were used. For this, assessments from economically similar regions were considered and merged with any available subregional estimates. For example, estimates from WPR A were merged with those from AMR A, EMR A and EUR A.

Global proxies

If these economic proxies were still not sufficient to obtain the required number of experts contributing to the assessments, then global proxies were employed, i.e., all experts' assessments for a given hazard were merged to provide global attribution estimates.

2.10. Software

We used Qualtrics software to collect expert information and assessments, a GDPR compliant tool. Qualtrics also enabled the automatic tailoring of the calibration questions and questions of interest based on experts' hazard/sub-region selection. The survey allowed the implementation of conditions for the input, such as increasing percentiles or non-numerical inputs. Moreover, it also allowed for implementing thresholds on the sums of median estimates across pathways and foods, and experts received messages to indicate exceptionally low/high median sum values, as detailed in Section 2.7.

The implementation of all the questions, for all hazards/subregions led to more than 9000 questions in total. Qualtrics could not export the collected expert data efficiently for such a large survey, separate links were generated for each hazard.

The Qualtrics link offered a flexible working tool, where experts could return to the survey at any time without losing already provided data. However, if experts' PCs had a firewall installed, data held on the PC could be deleted if the link to the Qualtrics platform was lost or closed.

Expert data were exported from the Qualtrics platform by the analysts and formatted for further analysis in R (version 4.4.2). Assessments for calibration questions and questions of interest were stored separately for each individual expert, who received an anonymized id, and CSV files for each hazard/subregion were generated, following the format reported in Appendix E (containing expert anonymized data). The R code for the Classical Model (Colonna et al., 2022) was adapted for the analysis. The Classical Model was applied for each hazard/subregion combination, except when expert data were merged for obtaining regional/economic/global proxies as described above. Experts' distributions were aggregated using five weighting schemes: equal, global, global optimal, item and item optimal and the best performing DM for the calibration questions was selected. 10,000 samples from pathways and food groups distributions were drawn and normalized so that the pathways' means and the food group means summed to 100%. This normalization was done for each sample.

2.11. Ethics and declaration of interest

Both experts and elicitors provided their consent to participate in the study. They were informed about the purpose of the study, the procedure and duration of the study, as well as how their data will be used, confidentiality and their rights. Thereafter, they provided a declaration of informed voluntary consent to participate as an elicitor or expert in the study. They were informed that they had the right to withdraw from the study at any time. Furthermore, the experts were also asked to

provide a declaration of interest, to disclose any circumstance that could represent a potential conflict of interest. Finally, faculty data steward and TU Delft's human research ethics committee has evaluated that there are no personally identifiable research data (PIRD) used for the source attribution results and that a formal approval is therefore not needed.

2.12. Expert and elicitor feedback

Both experts and elicitors were offered the opportunity to provide feedback on the elicitation process. Experts were asked whether the purpose of the study and questions were clear, if they found the training materials helpful, and felt they received sufficient information and support from the study team. The elicitors were asked how long the interview took and whether the questions seemed clear to the expert, and whether the experts seemed to have a good understanding of the elicitation protocol, required steps and the assessments requirements.

3. Results

This section summarizes the collected expert data, presents performance scores for the experts and metrics for the resulting Classical Model Decision Makers (DMs) solutions (defined in 2.8). As such, these evaluations offer an overall way to gauge the focal backing elicitation results provided for the 2026 WHO source attribution study. Finally, source attribution results are depicted.

3.1. Expert data

Over 850 respondents were registered to WHO's open call for experts, from which 290 were identified as eligible experts for the broad classes of hazards and invited to contribute to the study. From the experts approached through WHO/FERG networks and conferences, 119 were further engaged by the study team. In total, 379 experts were contacted to contribute to the study. From those, 196 responded positively and received their elicitations. Finally, assessments on target questions from 160 experts were used in our analysis. An additional expert provided only assessments for major pathways. The list of all experts contributing to the target questions is given in Table C2 in Appendix C. All expert input data are reported on Github (WorldHealthOrganization/FERG_2025).

The survey link with the detailed qualitative self-reported expertise was accessed over 400 times. The training materials for the experts were accessed 476 times by 203 experts, and those for the elicitors were accessed 126 times by 50 elicitors. Approximately 70% of the experts accessing the training materials watched the videos and answered the training questions. The feedback form was accessed 51 times by the experts and 21 times by elicitors. Insights into the feedback will be provided elsewhere. Further analysis of this data will follow in a subsequent manuscript.

Most of the online, one-to-one elicitations were conducted in English. Around 15 elicitations were also conducted in Spanish, Portuguese, Mandarin, Japanese and French. During 3 elicitations, experts preferred being accompanied by a colleague, to support with translation.

Most experts provided their source attribution assessments via Qualtrics. For around 40 experts, Excel files were prepared for each hazard and shared via email. This was especially efficient when experts provided assessments for all subregions for a given hazard.

A list with the elicitors who conducted at least one elicitation is included in Table C1 in Appendix C.

During the elicitation, some experts selected only a subset of the hazards initially listed in the Qualtrics expertise survey, whereas others did not provide assessments for (all) the selected hazards during the elicitation. Following the elicitor-led consultation, and as the self-administered work on target items progressed, some experts became unresponsive to follow-up approaches from the elicitation team; or they couldn't find the time to complete their assessments. Other experts mentioned that, as they gained

a better understanding of (some of) the target items, they felt no longer able to contribute judgments on those hazards. This led to a reduction in assessments for some hazards/sub-region combinations and, in some cases, to data gaps. Additional substantial efforts were engaged by WHO to cover those gaps. Despite these efforts, some hazards/sub-regions combination remained scarcely populated with expert assessments. Imputation strategies were then developed to address those challenges, as detailed in Section 2.9.

Nevertheless, the present study is scaled up considerably, relative to the previous WHO 2015 assessment. Table 3.1 compares the core parameters of the two studies. Since both studies used the Classical Model, the comparison will be straightforward.

Table 3.1 Scope and scale of the WHO 2026 study compared with the WHO 2015 study. Number of experts, hazards, subregions, distinct panels and individual expert scores.

	Experts	**Hazards**	**Subregions**	**Distinct panels**	**Individual expert scores**
2015	72	24	15	134	1675
2026	160	40	17	505	3237

In the WHO 2015 study, the calibration variables were the same for each expert panel. For the current study, in contrast, calibration variables were panel specific, as detailed in Section 2.2. Many experts participated in several panels and thus their individual information and statistical accuracy scores could differ slightly from one panel to another. This is why, in Table 3.1, the number of expert scores is much greater than the number of experts.

3.2. Elicitation process and data imputation

Experts provided attribution estimates for selected hazard/subregion. The number of expert assessments per subregion is included in Figure 3.1.

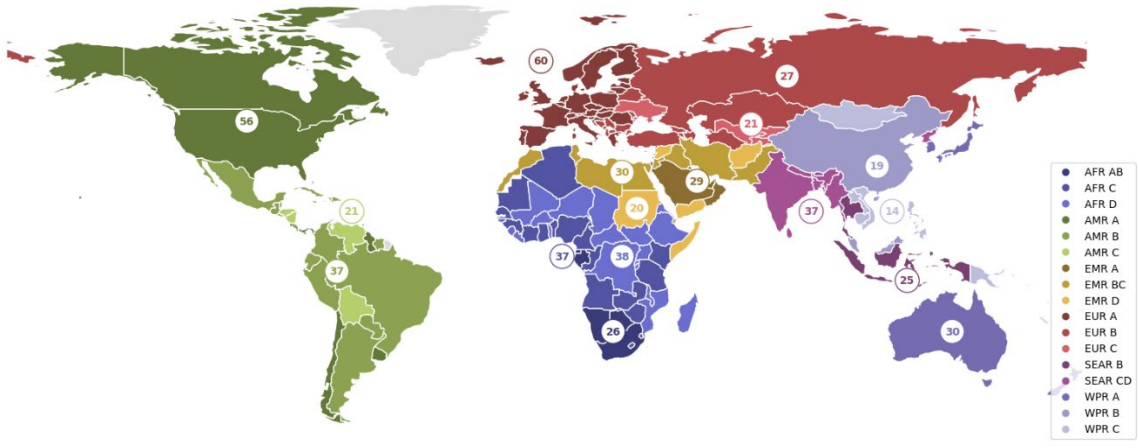

Figure 3.1. World map with the 17 subregions and the number of expert assessments per subregion.

The subregions highest number of expert assessments were EUR A and AMR A, whereas WPR B and C registered the lowest number of expert assessments.

Panels were formed for each hazard/subregion combination. For *E. multilocularis* and *Trypanosoma cruzi*, only endemic subregions were considered. For *Angiostrongylus* and *Sarcocystis*, very limited expert data were collected, resulting in estimates only for EUR subregions and SEAR B (for

Angiostrongylus). This led to a total of 632 panels for which source attribution estimates were obtained. As mentioned before, some panels did not meet the number of experts requirements and regional, economical or global proxies were employed. When proxies were employed, for example in two neighbouring subregions, and assessments in these subregions were merged, this resulted in duplicated source attribution results for the two subregions. This resulted in 505 distinct panels, as reported in Table 3.1.

The number of expert assessments per panel is included in Table 3.2-3.5. Green shading denotes panels for which specific subregional estimates have been obtained with at least 4 experts providing subregional assessments. Orange represents the panels for which regional proxies were employed. In purple, subregions for which economic proxies were imputed and in blue, subregions for which global proxies are highlighted. Overall, AMR and EUR subregions were most represented, whereas WPR subregions were least represented. A notable exception for the WRP subregions is exhibited for the chemical hazards, due to an intense effort from WHO to engage with domain experts. In terms of broad hazard classes, enteric (diarrhoeal) were mostly assessed, whereas parasites were least assessed. Engaging domain experts was particularly difficult for parasites, except for the EUR subregions. Overall, 213 panels gathered sufficient expert data for tailored subregional estimates, which amounts in almost 34% of all panels. In almost 56% of the panels, regional proxies were employed. Only 2 panels required global proxies, whereas economic proxies were employed for 64 panel (approximately 10% of the analysed panels).

Table 3.2. Counts of experts providing assessments for diarrhoeal enteric hazards /subregion combinations. In green, panels for which specific subregional estimates have been obtained (with at least 4 experts providing assessments). Orange cells represent panels for which regional proxies were employed. In purple, economic proxies were imputed to the panel and in blue, global proxies were applied. The abbreviations are explained in Table 1 in Appendix A.

	AFR AB	AFR C	AFR D	AMR A	AMR B	AMR C	SEAR B	SEAR CD	EUR A	EUR B	EUR C	EMR A	EMR BC	EMR D	WPR A	WPR B	WPR C	ALL
CAMP	5	5	3	18	8	4	5	5	22	4	3	6	4	1	5	3	2	103
CRYP	3	4	3	11	4	4	4	5	12	7	3	2	3	2	3	2	2	74
CYCL	1	2	1	9	3	3	2	2	4	5	2	1	2	1	2	1	1	42
ENTA	4	3	3	5	4	4	3	5	4	5	3	2	3	2	2	2	2	56
EAEC	1	4	1	5	2	1	2	3	4	1	1	3	3	0	1	1	0	33
EPEC	4	7	5	9	3	1	4	4	5	1	0	3	5	0	2	2	0	55
ETEC	2	6	2	6	5	2	4	4	5	1	0	3	5	0	2	2	0	49
GIAR	3	4	3	6	3	3	4	2	9	5	3	3	4	2	4	2	2	62
NORO	4	4	4	15	5	5	5	4	13	5	3	6	5	3	8	4	3	96
ROTA	3	4	3	6	2	2	2	2	7	2	1	3	2	1	2	2	1	45
NTS	4	5	4	15	8	2	5	4	12	4	2	4	5	1	7	3	2	87
SHIG	1	3	2	10	1	1	1	2	6	0	0	1	3	0	0	1	0	32
STEC	2	4	3	14	4	2	2	2	12	2	1	5	4	0	3	2	0	62
VIBR	2	3	3	7	2	2	4	2	4	1	1	4	5	1	2	2	1	46

Table 3.3. Counts of experts providing assessments for non-diarrhoeal enteric hazard/subregion combinations. In green, panels for which specific subregional estimates have been obtained (with at least 4 experts providing assessments). Orange cells represent the panels for which regional proxies were employed. In purple, economic proxies were imputed to the panel and in blue, global proxies were applied. The abbreviations are explained in Table 1 in Appendix A.

	AFR AB	AFR C	AFR D	AMR A	AMR B	AMR C	SEAR B	SEAR CD	EUR A	EUR B	EUR C	EMR A	EMR BC	EMR D	WPR A	WPR B	WPR C	ALL
BRUC	3	1	3	7	3	1	2	2	5	4	1	2	3	1	2	0	0	40
CPERF	1	2	3	11	6	3	2	2	7	4	1	5	4	1	3	1	1	57
Cbot	2	2	2	9	4	2	3	4	7	3	2	5	3	1	3	1	1	54
HAV	2	2	1	11	3	2	1	1	6	1	1	6	6	2	4	2	1	52
LIST	3	4	1	16	6	3	3	4	15	4	2	6	4	1	4	3	1	80
zMTBC	3	5	5	5	3	0	0	2	4	2	0	1	1	0	1	0	0	32
PARA	3	2	2	5	3	1	1	1	3	2	0	3	4	0	1	0	0	31
TYPH	4	4	5	7	3	1	3	4	4	2	0	7	5	0	1	0	0	50
HEV	0	0	0	3	2	2	0	2	9	2	1	2	2	0	2	1	0	28
STAPH	2	6	4	11	8	3	4	6	8	3	1	5	4	1	1	1	0	68

Table 3.4. Counts of experts providing assessments for parasitic hazard/subregion combination. In green, panels for which specific subregional estimates have been obtained (with at least 4 experts providing assessments). Orange cells represent panels for which regional proxies were employed. In purple, economic proxies were imputed to the panel and in blue, global proxies were applied. The abbreviations are explained in Table 1 in Appendix A. For subregions in grey,, no attribution results are produced.

	AFR AB	AFR C	AFR D	AMR A	AMR B	AMR C	SEAR B	SEAR CD	EUR A	EUR B	EUR C	EMR A	EMR BC	EMR D	WPR A	WPR B	WPR C	ALL
ASC	4	4	3	3	3	3	3	4	6	6	3	2	3	2	2	2	2	55
Emult	1	1	1	4	2	2	2	3	9	6	3	1	2	2	2	2	2	45
Egan	3	3	3	4	4	3	3	4	7	6	2	2	2	2	2	2	2	54

FASC	1	2	2	1	1	1	1	1	5	4	2	1	1	1	1	1	1	27
TOXO	3	3	2	5	3	3	3	3	10	5	2	2	2	2	3	1	1	53
TRICH	1	1	1	8	3	2	2	2	7	5	2	2	1	1	1	1	1	41
Tcru	1	1	1	3	6	4	1	1	2	3	2	1	1	1	1	1	1	31
TOXOC	1	1	1	1	2	2	1	1	4	4	2	1	1	1	1	1	1	26
ANGIO	0	0	0	0	0	0	1	0	3	3	1	0	0	0	0	0	0	8
SARCO	0	0	0	0	0	0	0	0	2	3	1	0	0	0	0	0	0	6

Table 3.5. Counts of experts providing assessments for chemical hazards/subregions combination. In green, panels for which specific subregional estimates have been obtained (with at least 4 experts providing assessments). Orange cells represent panels for which regional proxies were employed. In purple, economic proxies were imputed to the panel and in blue, global proxies were applied. The abbreviations are explained in Table 1 in Appendix A.

	AFR AB	**AFR C**	**AFR D**	**AMR A**	**AMR B**	**AMR C**	**SEAR B**	**SEAR CD**	**EUR A**	**EUR B**	**EUR C**	**EMR A**	**EMR BC**	**EMR D**	**WPR A**	**WPR B**	**WPR C**	**ALL**
AB1	5	6	8	4	2	2	4	6	7	5	3	3	5	3	3	6	2	74
iAs	3	4	3	6	1	1	2	6	4	3	2	3	3	1	4	4	2	52
Cd	2	3	2	6	1	1	2	4	4	4	2	1	4	1	7	5	2	51
DIOX	3	4	4	5	1	1	2	3	7	4	2	1	3	3	6	3	1	53
Pb	2	2	1	5	1	1	3	7	5	2	2	2	4	3	5	7	2	54
MeHg	3	4	3	5	4	3	3	6	5	4	2	3	4	2	8	6	2	67

As mentioned above, for subregions that required proxies, experts' assessments were merged from neighbouring subregions, from similar economic subregions or from global analogues. In most of the cases, assessments from two neighbouring subregions were merged and the corresponding results were reported for both subregions. For subsequent analysis purposes, only one panel was retained; this led to 505 distinct panels being analysed in this study.

Figure 3.2 shows the spread of panel sizes for this study, ranging from 4 experts to 22 per panel (one global proxy panel had assessments from 27, not shown); it is noteworthy that smaller panels dominate the count distribution.

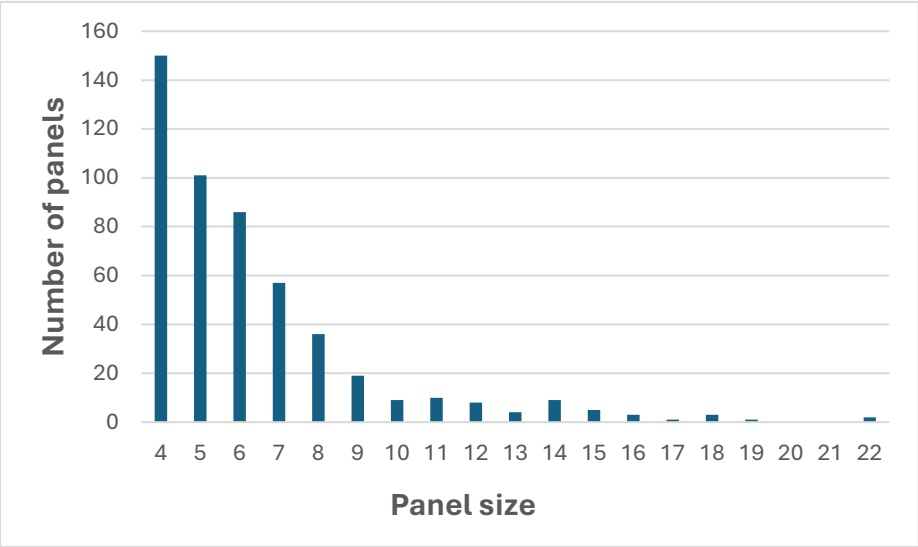

Figure 3.2 Distribution of panel sizes, by number of participating experts.

We see that 150 of the 505 panels had only four experts and 449 had less than ten. This shows that despite the merging of some assessments, the number of assessments per panel remained limited. This is also reflected in the somewhat erratic behaviour of DM performance-based measures. Distributions of panel sizes, by hazard and by subregion, are shown in Figure 3.3.

In terms of numbers of experts engaged, panel sizes for diarrhoeal enteric hazards are the most varied, while the chemical panels are relatively uniform in size, albeit at quite small numbers of experts, generally less than ten persons per panel.

For the different regions, Figure 3.3 indicates WPR subregions, EMR and AFR C have smaller panels when addressing the series of hazards, than most other subregions.

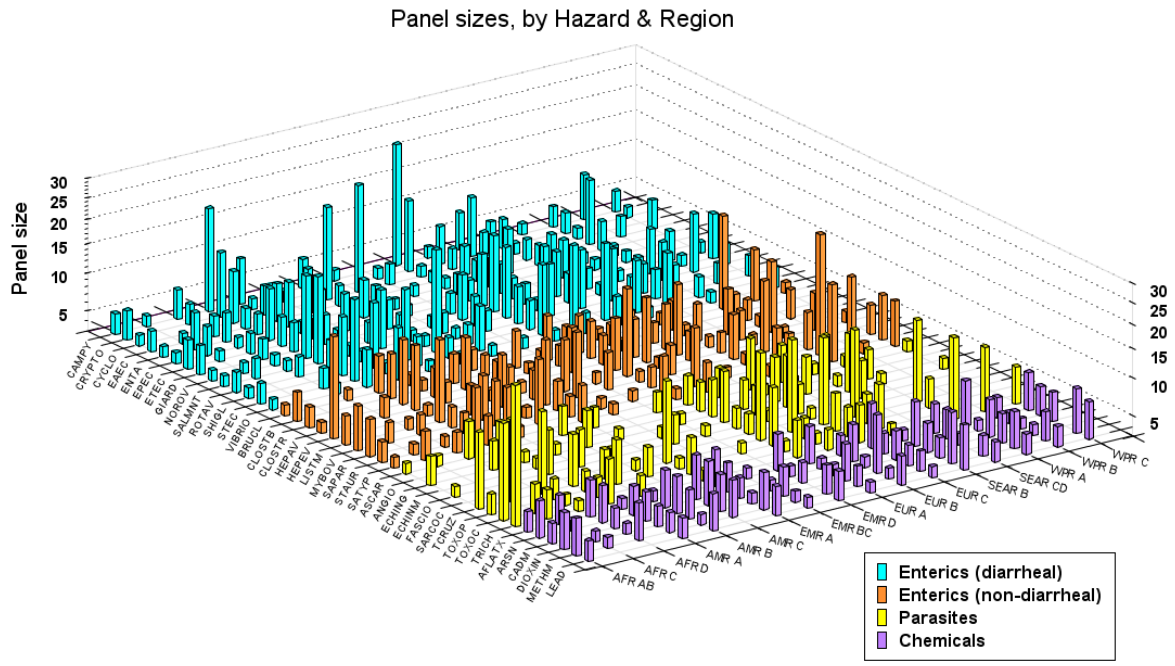

Figure 3.3 Panel sizes, by hazard and by subregion. Hazard types are colour-coded per category: diarrhoeal enteric hazards; non-diarrhoeal enteric hazards; parasitic, and chemical hazards. Abbreviation codes on the plot, for hazards and subregions, are listed in full in Table 1 in Appendix A.

3.2. Expert assessments performance

Expert assessments have been evaluated using the two objective measures: Statistical Accuracy (SA) and informativeness, based on the tailored calibration questions. Table 3.6 shows the percentages of experts whose SA score fell below various thresholds. In the present study, 11% of experts' SA scores were above the traditional $p = 0.05$ threshold for simple hypothesis rejection, whereas in the WHO 2015 exercise that total was 6% of experts. For comparison, a review in 2021 (Cooke et al., 2021) found that 140 of 530 experts (26%) had SA scores above 5%. In general, the percentages of experts with low SA scores were greater in the WHO 2015 (Hald et al., 2016) study than they are in the present (WHO 2026) assessment. This points to a positive effect of training.

Table 3. 6 Percentages of WHO 2026 experts whose Statistical Accuracy (SA) fall below a certain threshold (columns 1 and 3) given hypothesis rejection p-values and comparative percentages from WHO 2015 study.

2026 % experts SA <=		2015 % experts SA <=	
0.05	89.38	0.05	94.44
0.01	75.02	0.01	87.50
0.001	57.23	0.001	75.00
0.00001	33.54	0.00001	41.67
1E-07	10.99	1E-07	13.89

Figure 3.4 displays the relation between experts' SA scores and their corresponding information (Inf(cal)) scores for calibration variables. There is a negative rank correlation between the two scores of -0.44. This shows that high informative scores are associated with lower statistical accuracy scores.

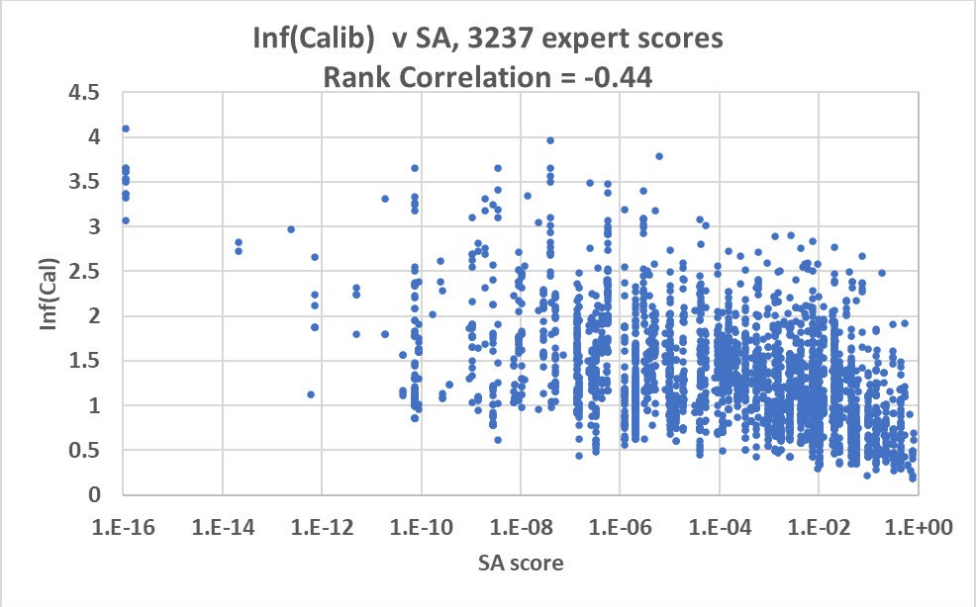

Figure 3.4 Experts' information scores for the calibration questions [Inf(cal] as a function of their Statistical Accuracy [SA] scores.

This negative correlation is further illustrated in Figures 3.5a and 3.5b, showing the rolling rank correlations between Inf(cal) and SA for the WHO 2026 and WHO 2015 studies, respectively.

In the case of the present (WHO 2026) data, values plotted are average SA and Inf(cal) values for individual experts because, in this study, most experts completed multiple region- and hazard specific calibrations, with seed items that varied from one hazard-region combination to another. In the WHO 2015 assessment, one set of seed items was used to generate universal calibration scores for all scenarios, per expert.

In both Figures 3.5a and 3.5b, the plots position experts by their Inf(cal) rank correlation, where their SA (or average SA) exceeds the threshold value on the horizontal axis.

For the rest of the experts, sitting below these thresholds, the two plots are similar: a negative rank correlation of about -0.4 characterizes the pattern for most experts in both the WHO 2026 and WHO 2015 study cohorts.

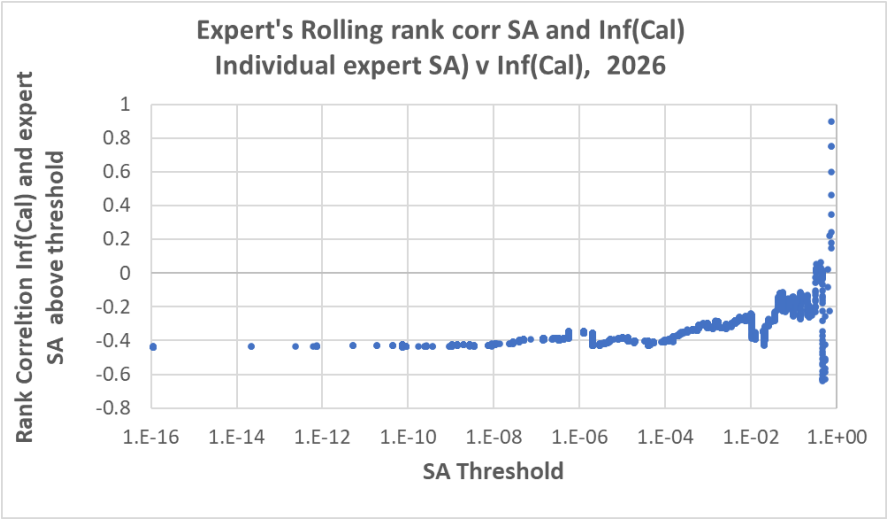

Figure 3.5a Rolling rank correlation, WHO 2026 elicitation; individual expert average Inf(cal) and average SA (above threshold).

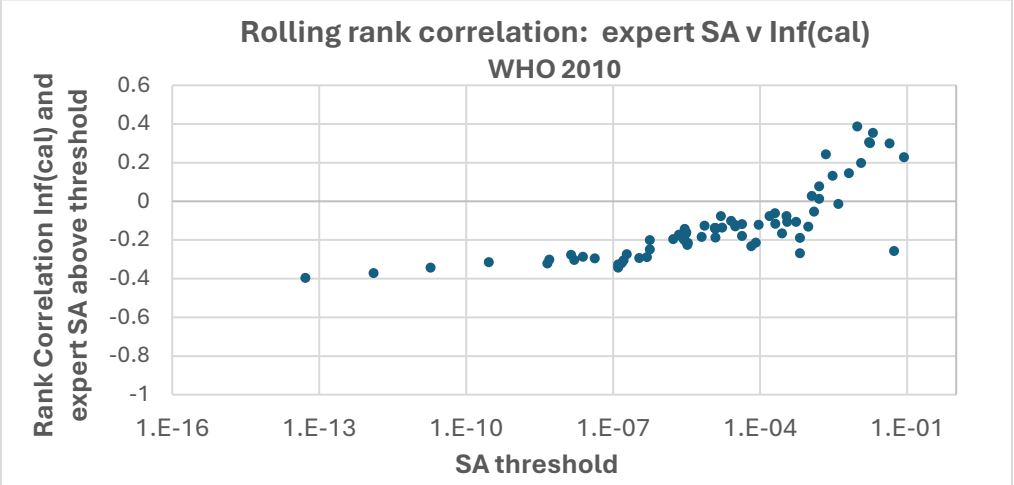

Figure 3.5b Rolling rank correlation, WHO 2015 elicitation; individual expert Inf(cal) and SA (above threshold).

For both studies, WHO 2026 and WHO 2015, the correlations become unstable for high SA values owing to the relatively small number of experts above the threshold. However, this volatility emerges at a higher SA threshold for the WHO 2026 data (i.e., above SA threshold 1E-2: see Figure 3.4a), than for the WHO 2015 case (i.e., above SA threshold 10^{-4}: see Figure 3.4b).

The sheer size of the present WHO 2026 elicitation – involving 160 experts participating in 505 panels generating a total of 3237 individual calibration scores (see Table 3.1, above) – offers an exceptional opportunity to provide an in-depth, mass contextualization of individual performance scores against performance scores for DM solutions. For this scrutiny, the DM solutions are those which are based on the preferred optimised Item Weights pooling.

Figures 3.6a, b present the same data – i.e., information scores and SA scores for individuals versus DMs – in two slightly different forms.

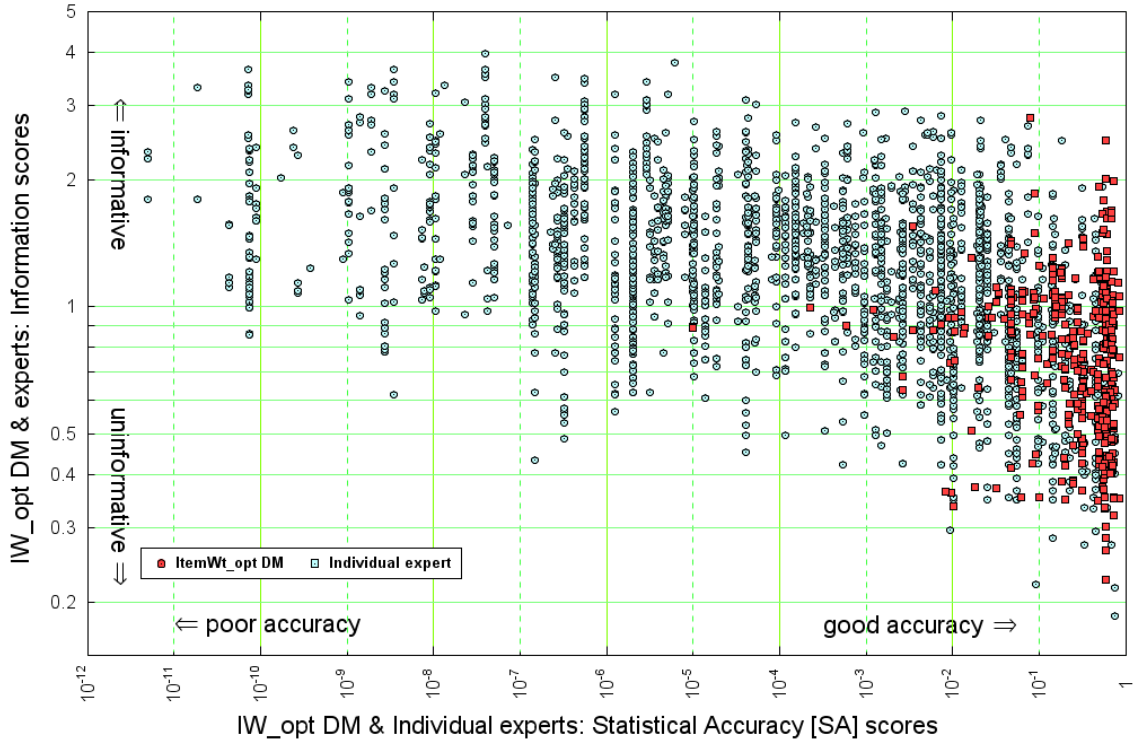

Figure 3.6a Comparison of IW_opt DM information scores versus individual experts' Information scores, as a function of Statistical Accuracy (SA) scores – present WHO 2026 study.

In Figure 3.6a, individual expert calibrations manifest information scores which, collectively, trend marginally higher than those of ItemWt_opt DM solutions.

However, in terms of statistical accuracy (SA) performance, DM scores cluster in the range SA ~ 5 x 10^{-2} to SA ~ 1, with the majority toward the upper end of this range. (Recall, $p_0 > 0.05$ is a traditional threshold for null hypothesis non-rejection).

In contrast, the bulk of individual expert scores are massed at much lower SA scores than the DMs, most with SA < 5 x 10^{-2}, extending down to SA ~ 10^{-16} (note that x-axis is log-scale, and censored at SA = 10^{-12}).

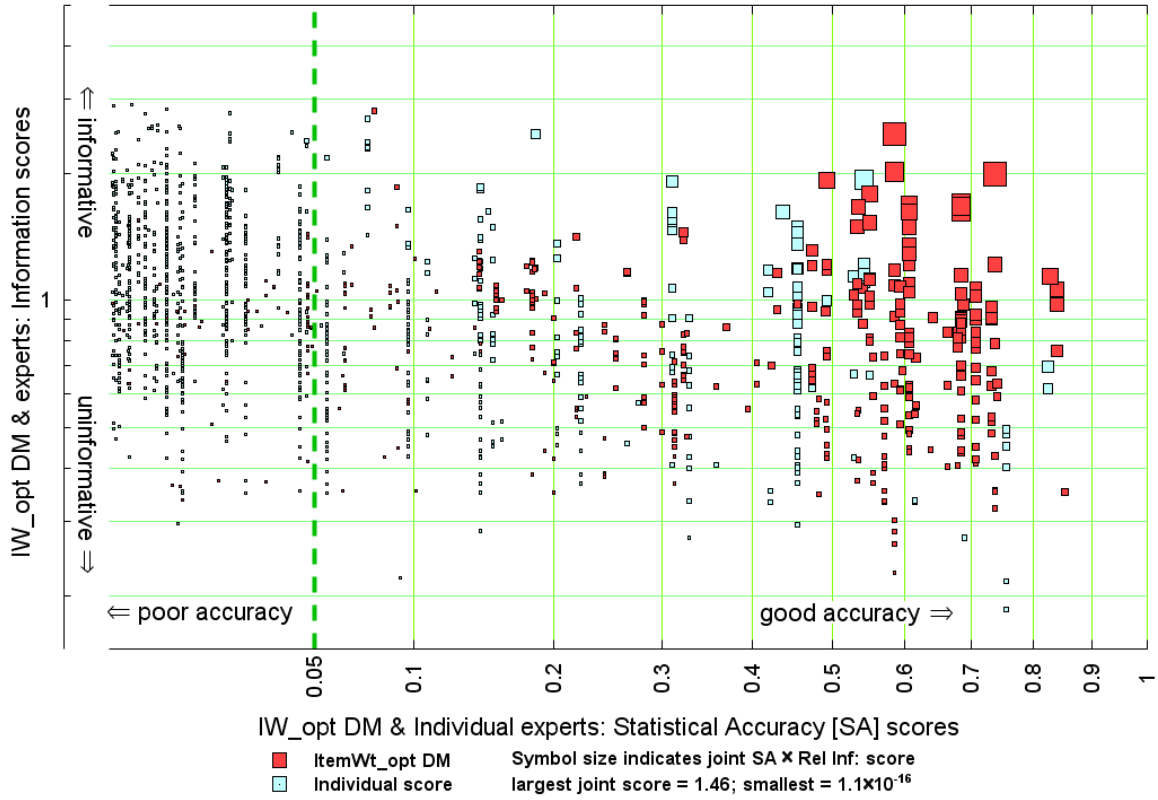

Figure 3.6b Data from Figure 3.5a plotted against Statistical Accuracy SA, with the x-axis re-scaled to better expose Information scores for solutions with higher SA values. In this plot, sizes of symbols for IW_opt DM and for individual experts reflect their respective combined scores. A green dashed line highlights the 0.05 threshold.

Figure 3.6b shows the same data as Figure 3.6a but reframed with a square root x-axis to better reveal details of data points with high SA scores. Also in this plot, the sizes of data points are scaled in proportion to the corresponding Combined (unnormalized) weights (i.e., SA x calibration Relative Information scores). The larger the symbol the stronger the performance of the individual judgment or the DM solution and hence the weight that can be ascribed to its implications.

Clearly, a few individual experts perform well in these terms and, indeed, these are few in number. Moreover, there is no way – *a priori* and in the absence of calibration – of knowing who these high scoring experts are. And, even when they are identified, it is very usual to find that optimal DM solutions often meld in judgments from other experts with lower performance metrics.

All this helps emphasize that, in the present study, IW_opt DM scores, and the resulting pooled solutions, are predominantly superior to judgments from individual experts. Given the magnitude of the present study, and hence the uniquely substantive nature of this evidence, key features of DM performance warrant further analysis, as follows.

3.3. Decision Makers' Performance

Five Decision Makers (DMs) were computed for each of the 505 panels in WHO 2026. Table 3.7 shows the mean scores and standard deviations. It is notable that Equal Weights EWDM scores a bit higher in SA than the other DMs. This reflects the above noted erratic behaviour of the performance-based measures. Nonetheless, EWDM is the least informative DM, which is then reflected on the lowest combined scores. Also notable is the minimal difference between average combined scores for the optimized DMs, as well as for the IW DM.

Table 3.7 Summary of five DMs scores from WHO 2026 data, with respect to statistical accuracy (SA), and informativeness (Inf), calculated for the calibration questions (Inf(cal)) and calibration and target questions (Inf(all)), and combined score (Comb). For each DM, the average scores (ave) and standard deviation (stdev) are presented.

Score	SA	Inf(all)	Inf(cal)	Comb
DM	**PW**			
ave	0.33	1.70	0.64	0.19
stdev	0.24	1.07	0.32	0.17
DM	**EW**			
ave	0.41	1.20	0.39	0.16
stdev	0.22	0.56	0.16	0.11
DM	**PW_opt**			
ave	0.39	2.10	0.73	0.27
stdev	0.24	1.30	0.37	0.23
DM	**IW**			
ave	0.31	2.24	0.74	0.22
stdev	0.23	1.07	0.27	0.20
DM	**IW_opt**			
ave	0.35	2.45	0.82	0.27
stdev	0.25	1.32	0.34	0.24

Table 3.8 shows the percentages of WHO 2026 panels in which the various DM SA scores fall below a given *p*-level rejection threshold. Here again we see that EW has fewer very low scores and more scores above $p = 0.1$. From this, we infer that the pre-elicitation training only partially compensated for the small size of some panels.

Table 3.8 Percentage of panels for which DMs SA scores are beneath p thresholds, for the five DMs.

p threshold	PW	EW	PW_opt	IW	IW_opt
0.1	23.37	13.66	16.04	27.92	23.56
0.05	18.61	7.92	11.29	16.04	13.86
0.01	5.15	1.19	2.57	4.75	3.96
0.001	1.19	0.00	0.79	1.19	1.19
0.0001	0.00	0.00	0.00	0.59	0.59
1.0E-05	0.00	0.00	0.00	0.59	0.59
1.0E-06	0.00	0.00	0.00	0.59	0.00

The combined score rewards both statistical accuracy and informativeness and is the best overall performance measure (see Section 2.8). The DM with the highest Combined score is chosen as the rational SEJ consensus for each hazard/subregion panel in the present study. Table 3.9 shows that, when

it comes to combined scores, the two DM optimized performance scores, PW_opt and IW_opt (see Section 2.8), are overwhelmingly superior SEJ pooling solutions, relative to others.

Parenthetically, because PW_opt and IW_opt solutions may coincide when one unique expert receives weight 1 in both forms of optimized DM solution, there are several tied maximal scores: while there are 505 panels, the reported maximal score counts in Table 3.9 sum to 614.

Table 3.9 Number of WHO 2026 panels in which DM's had maximal combined scores, from the 505 panels; for 109 panels, ties were recorded.

PW	PW_opt	IW	IW_opt	EW
55	209	39	235	76

IW_opt (i.e., optimized Item weights DM solutions) emerge as the most frequently preferred option for characterizing most hazard/subregion panel expert judgments (in cases where the combined score of another DM is higher, that DM is chosen). Figure 3.7 breaks down the overall performance of EW and IW_opt into SA and Inf(Cal) scores, ordered by IW_opt SA scores. In 52% of the 505 panels, the IW_opt SA score is greater or equal to that of EW. The combined score for each expert, when normalized over all panel members, sums to 1 and is strongly dominated by SA.

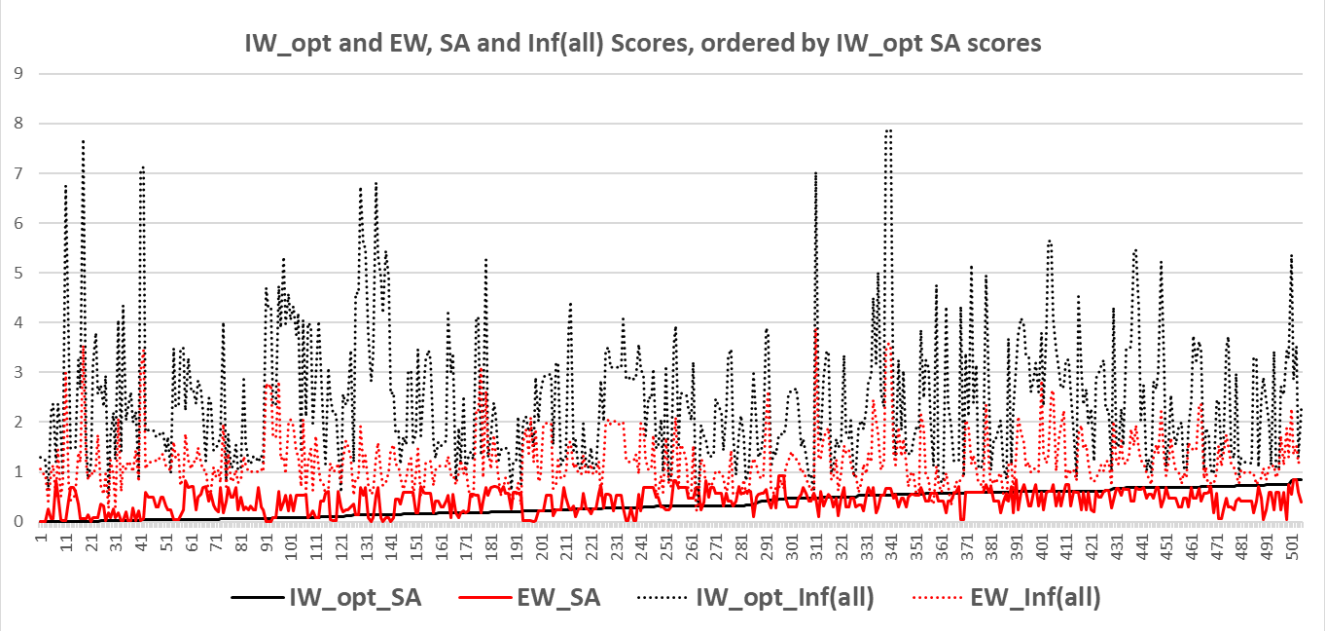

Figure 3.7 IW_opt and EW scores for SA and Inf(all) for the 505 panels. The scores are ordered by the IW_opt SA scores.

Figure 3.8 plots the combined scores of EW and IW_opt as functions of EW_SA. We see that IW_opt's gain in information compensates for modest differences in SA scores and that the differences in combined scores increase as EW_SA increases. In 69% of the 505 panels in the WHO 2026 study, the combined score of IW_opt is greater than or equal to that of EW. For context, a factor 2 increase in Information score corresponds roughly to halving the width of the 90% central confidence band.

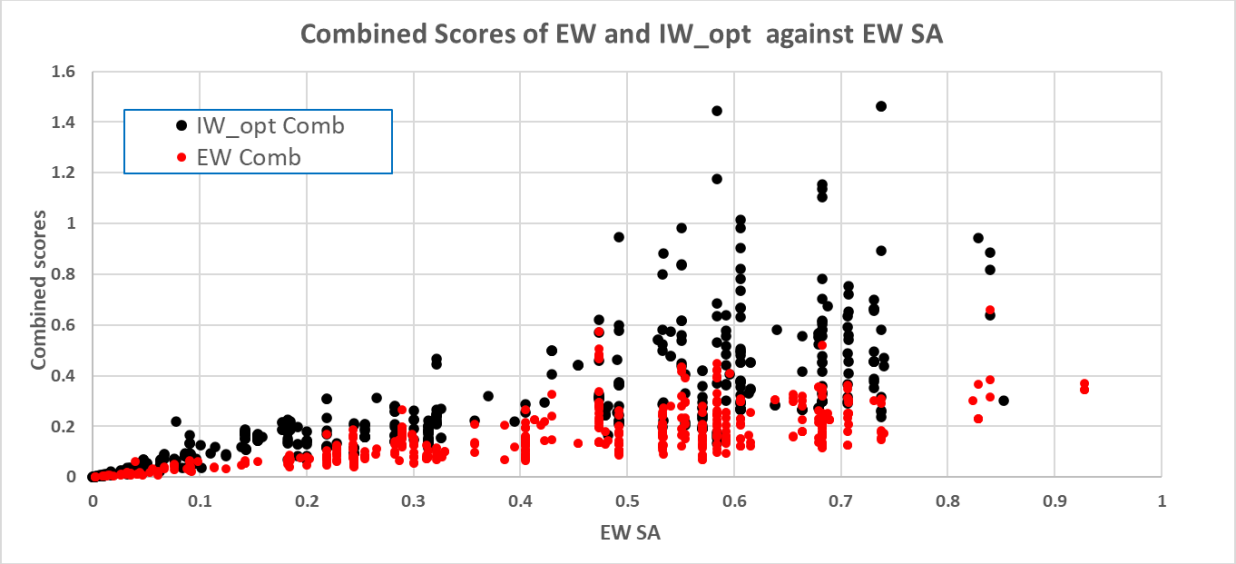

Figure 3.8 Combined scores of EW and IW_opt against EW SA.

Figure 3.9 plots SA scores for EW and IW_opt as a function of panel size. For small panel sizes, very low IW_opt SA scores are slightly more abundant than for EW. Of the 505 panels, EW has one SA score below 10^{-3}, IW_opt has five.

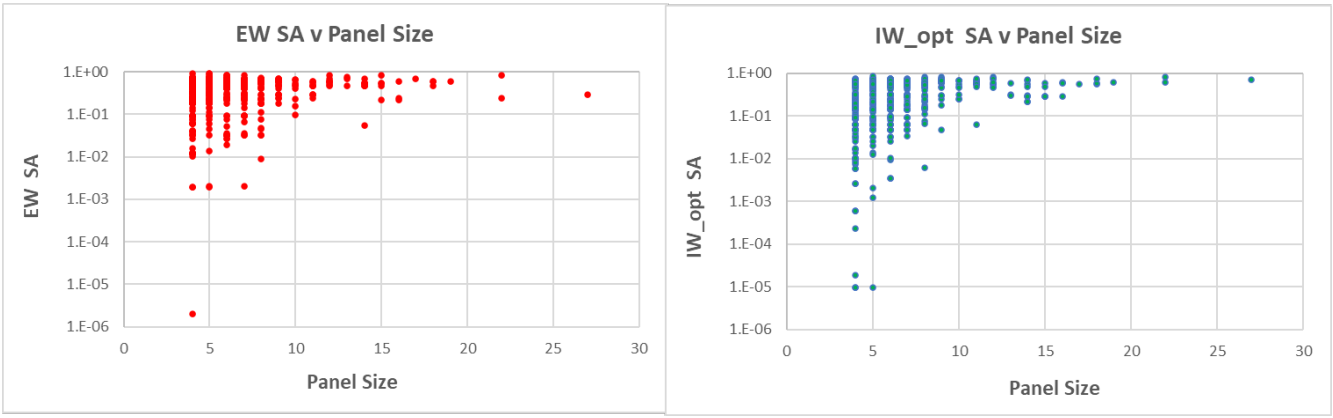

Figure 3.9 EW SA (left) and IW_opt SA (right) against panel size.

The study-wide generic statistics just presented - on Statistical Accuracy and Information performances for IW_opt and EW decision makers - leads to three mappings of comparative performances of the 505 expert panels in terms of hazard type and subregion, Figures 3.10– 11. On these plots, hazard are color-coded per category: diarrhoeal and non-diarrheal enteric hazards; parasitic, and chemical hazards – see Table A2 – where abbreviation codes for hazards and subregions are also listed.

The first of the three, Figure 3.10a, summarizes Statistical Accuracy (SA) scores per expert panel, with the upper frame of the plot showing cases where IW_opt SA score is greater than EW SA score. The lower frame identifies cases where the reverse is true.

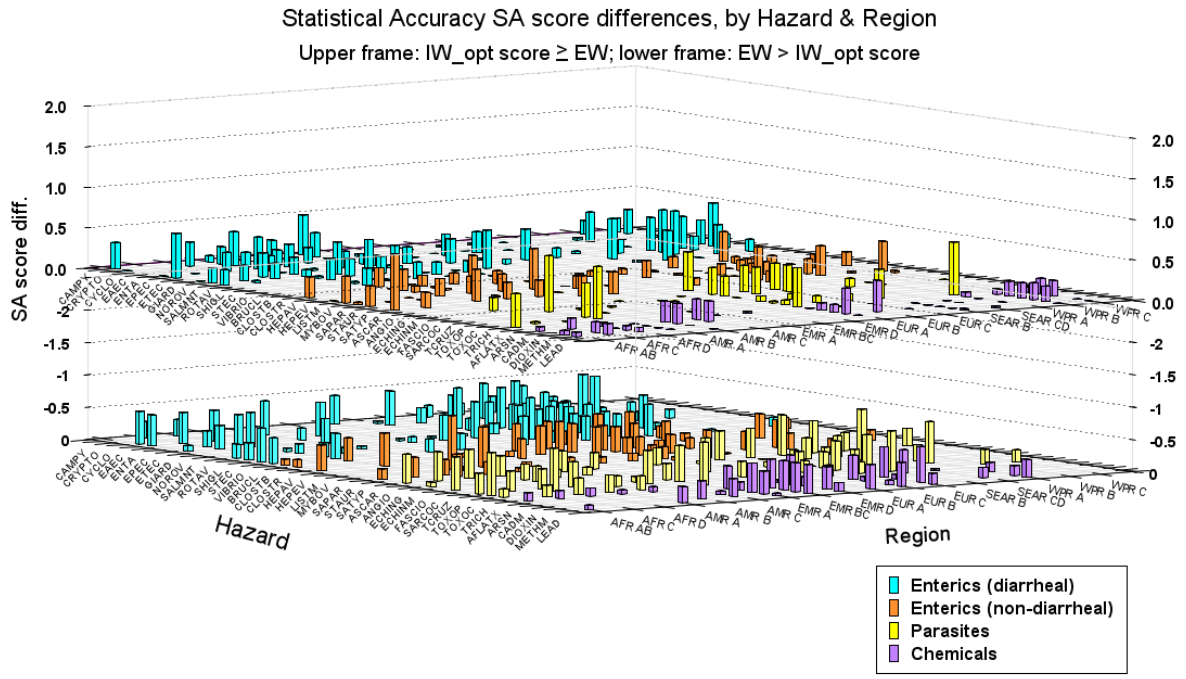

Figure 3.10a IW_opt DM and EW DM Statistical Accuracy (SA) scores in relation to Hazards and Regions.

Next, Figure 3.10b provides a similar vista, but for information score Inf(cal). For this DM metric, the picture is in stark contrast to that of Figure 3.10a. Here, IW_opt Inf(cal) scores are comprehensively better than EW Inf(cal) scores, in terms of numbers of cases and scales of differences.

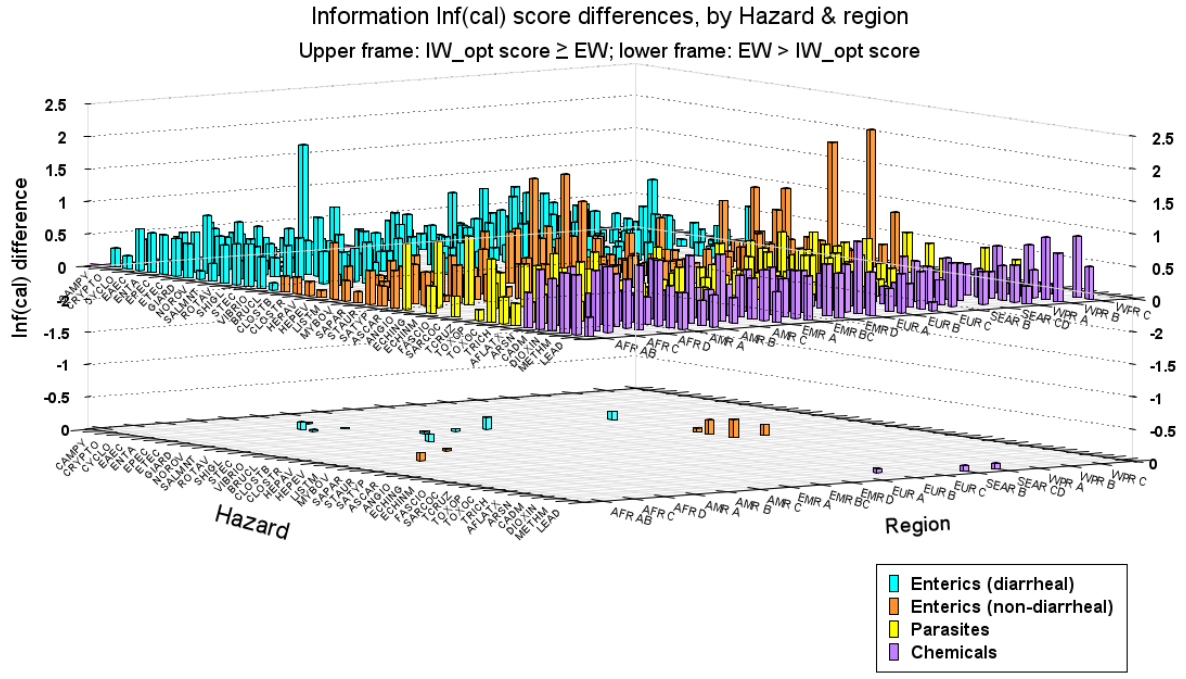

Figure 3.10b IW_opt DM and EW DM Information Inf(cal) scores in relation to hazards and subregions.

Figure 3.11 illustrates how the Classical Model DM un-normalized weights (i.e., combined scores) for each of these DMs vary across the wider scope of the WHO 2026 study. The two plot frames express

differences between the IW_opt Comb. score and that of the EW DM. The upper frame shows cases where the IW_opt Comb. score is greater than that of the EW solution for the same case, with bar height indicating how great that difference is. The lower frame identifies other cases, where the reverse is true.

Overall, the pattern is very clear: IW_opt DM is superior to EW DM in 69% of cases, and generally strongly so in terms of Comb. score differences. Generally speaking, differences (i.e., IW_opt DM combined score − EW DM combined score) are higher for disease hazards designated as diarrhoeal enteric hazards, declining through non- diarrhoeal enteric hazards, then parasitic, to chemical hazards. This said, there are relative variations within regions and sub-regions, for any given hazard type; these dissimilarities may be informative. For instance, the greatest single difference, in favour of IW_opt over EW, is +1.38 for diarrhoeal enteric hazards disease, CAMP (*Campylobacter_spp.*), and this in respect of expert panel judgments for region EMR A (WHO East Mediterranean A). In contrast, the greatest negative difference is only -0.33 for parasitic disease FASC (*Fasciola & Fasciolopsis*) in WHO Region SEAR CD (South-east Asia CD); most negative differences, plotted in the lower frame of Figure 3.11, are small.

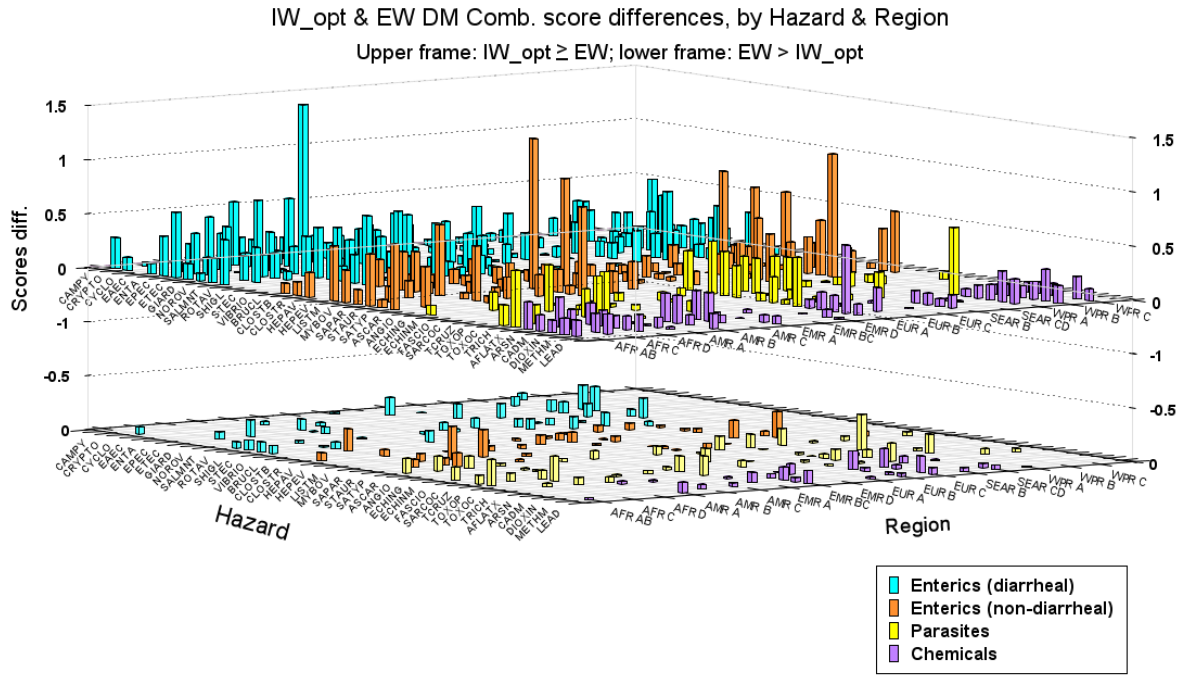

Figure 3.11 IW_opt and EW DM combined score differences, by hazard and by subregion.

In summary, Figures 3.9 - 11 show, in compact form, hazard/subregion DM-related performance details, determined from the larger and wider-ranging WHO 2026 elicitation study, in comparison with the earlier WHO 2015 assessment.

Table 3.10 shows that there is no strong effect of panel size or region on the panel average or standard deviation of EW SA or IW_opt SA. However, the correlation between panel size and EW SA is notably weaker than that with IW_opt. The latter DM is also somewhat more variable with respect to subregion than EW SA. Note that averaging over panel size or regions treats each panel size and subregion grouping as single samples, taking no account of the number of panels in each category.

We see that IW_opt SA is more sensitive to panel size than EW SA. Thus, the overall average SA for IW_opt from Table 3.3 is 0.35, whereas the average over panel size is 0.488. The overall average SA for EW is 0.412 whereas the average over panel size is 0.473. Both EW and IW_opt are degraded by smaller panel sizes, but IW_opt more so.

Table 3.10 Average SA per panel size and per subregion for EW and IW_opt. Average and standard deviation of SA scores and correlation with panel size.

Panel size	EW SA	IW_opt SA	Region	EW SA	IW_opt SA
4	0.388	0.274	AFR AB	0.458	0.399
5	0.409	0.329	AFR C	0.444	0.339
6	0.444	0.337	AFR D	0.396	0.243
7	0.376	0.335	AMR A	0.368	0.428
8	0.370	0.406	AMR B	0.385	0.373
9	0.473	0.536	EMR A	0.463	0.410
10	0.413	0.487	EMR BC	0.426	0.319
11	0.425	0.552	EMR D	0.454	0.247
12	0.599	0.623	EUR A	0.423	0.363
13	0.631	0.399	EUR B	0.407	0.276
14	0.477	0.415	EUR C	0.393	0.222
15	0.520	0.387	SEAR B	0.446	0.327
16	0.349	0.491	SEAR CD	0.453	0.277
17	0.679	0.554	WPR A	0.447	0.464
18	0.547	0.624	WPR B	0.310	0.397
19	0.584	0.606	WPR C	0.313	0.396
22	0.541	0.723			
27	0.289	0.707			
Standard deviation	**0.104**	**0.132**		**0.047**	**0.071**
Average	**0.473**	**0.488**		**0.412**	**0.343**
Correlation with panel size	**0.21**	**0.82**			

Table 3.11 summarizes various average DM scores for IW_opt and EW, by the forty different hazards.

Table 3.11 Average IW_opt and EW scores: statistical accuracy (SA), informativeness for calibration questions (Inf(cal)) and all questions (Inf(all)) and combined score (Comb), by hazard. Last column depicts the percentage gain, with respect to the combined score, of the IW_opt over EW. The abbreviations are explained in Table 1 in Appendix A.

Ave scores by Hazard	IW_opt SA	IW_opt Inf(all)	IW_opt Inf(cal)	IW_opt Comb	EW SA	EW Inf(all)	EW Inf(cal)	EW Comb.	IW_opt Comb %gain over EW
CAMP	0.474	2.122	0.837	0.420	0.514	0.921	0.386	0.202	108.3
CRYP	0.620	3.030	0.560	0.363	0.546	1.197	0.334	0.196	84.6
CYCL	0.365	4.623	0.943	0.336	0.617	2.070	0.408	0.243	38.2
EAEC	0.451	2.528	0.773	0.353	0.506	1.102	0.425	0.209	68.7
ENTA	0.628	4.102	0.797	0.511	0.577	1.855	0.399	0.223	129.2
EPEC	0.335	2.244	0.868	0.300	0.488	0.753	0.391	0.174	71.8
ETEC	0.298	2.094	0.741	0.192	0.468	0.824	0.340	0.156	23.2
GIAR	0.517	3.441	0.672	0.317	0.549	1.368	0.301	0.167	89.7
NORO	0.445	2.208	0.743	0.321	0.279	0.955	0.431	0.125	156.4
NST	0.456	1.138	0.818	0.381	0.498	0.567	0.403	0.196	94.0

ROTA	0.375	1.678	0.628	0.181	0.401	1.070	0.409	0.161	12.1
SHIG	0.429	1.585	1.081	0.476	0.458	0.840	0.566	0.266	78.8
STEC	0.434	1.939	0.715	0.305	0.517	0.945	0.396	0.199	53.7
VIBR	0.370	1.683	0.643	0.282	0.571	0.842	0.341	0.178	57.9
BRUC	0.384	2.665	0.885	0.352	0.457	1.407	0.361	0.157	123.6
Cbot	0.428	1.836	0.697	0.254	0.506	0.917	0.309	0.159	60.0
CPERF	0.414	2.149	0.607	0.252	0.389	1.254	0.341	0.137	83.4
HAV	0.338	2.595	0.712	0.239	0.569	1.117	0.368	0.195	22.2
HEV	0.424	1.959	0.573	0.241	0.454	1.264	0.359	0.124	94.8
LIST	0.535	2.075	0.816	0.451	0.494	0.888	0.334	0.167	169.7
zMTBC	0.494	1.102	0.594	0.257	0.631	0.730	0.413	0.261	-1.3
PARA	0.280	2.952	1.083	0.242	0.352	1.162	0.424	0.143	69.7
STAPH	0.370	2.089	0.772	0.309	0.337	1.170	0.406	0.136	126.7
TYPH	0.363	3.047	1.043	0.462	0.353	1.277	0.362	0.158	192.2
ASC	0.120	1.722	1.031	0.101	0.443	1.216	0.347	0.153	-33.5
ANGIO	0.062	1.924	0.851	0.054	0.128	1.350	0.366	0.049	10.3
Egan	0.347	1.438	0.810	0.267	0.349	1.083	0.395	0.134	99.3
Emult	0.437	1.252	0.755	0.326	0.342	1.038	0.414	0.154	112.4
FASC	0.215	2.861	0.784	0.141	0.350	1.392	0.323	0.119	18.8
SARCO	9.8E-06	1.241	0.895	1.241	0.002	0.966	0.290	0.001	-98.5
Tcru	0.304	2.893	0.588	0.190	0.494	1.853	0.337	0.166	14.2
TOXO	0.199	2.081	0.910	0.141	0.541	1.086	0.348	0.190	-25.7
TOXOC	0.326	2.671	0.579	0.200	0.202	1.539	0.302	0.063	216.9
TRICH	0.224	1.257	0.869	0.202	0.331	1.037	0.403	0.124	62.4
AFB1	0.198	5.258	1.013	0.176	0.242	2.779	0.520	0.112	57.5
iAs	0.167	2.743	0.959	0.156	0.194	1.024	0.425	0.082	89.7
Cd	0.158	3.211	0.963	0.148	0.159	1.356	0.464	0.067	120.3
DIOX	0.202	3.330	1.098	0.213	0.189	1.137	0.435	0.081	164.1
MeHg	0.162	3.239	1.117	0.184	0.291	1.900	0.391	0.110	68.1
Pb	0.207	1.793	0.815	0.146	0.356	0.687	0.391	0.141	3.3

The rightmost column records the % gain in Comb. score achieved by IW_opt, relative to the corresponding EW score.

The first thing to note is that the IW_opt Comb. score is more than double that of EW for ten Hazards (highlighted); the greatest gain, +216.9%, is achieved for TOXOC (Toxocara). Also in this column, there are just four negative percentage gains in the forty, i.e., zMTBC (*Mycobacterium bovis/caprae/orygis*); ASC (*Ascaris lumbricoides*); SARCO (*Sarcocystis*) and TOXO (*Toxoplasma gondii*). According to the tabulated averages, negative percentage gains indicate that, for the four identified hazards, in each instance the EW Comb. score is greater than the IW_opt Comb. score. Examination of the relevant datasets may reveal factors influencing these particular 'negative gain' results for the IW_opt DM SEJ pooling option.

In short, there are thirty-six hazards in Table 3.7 where IW_opt DM outperforms EW DM – often by large margins – at least based on these alternative hazard/averaged combined scores.

Table 3.12 mirrors Table 3.11, this time showing various average DM scores for IW_opt and EW, estimated for each of the seventeen subregions.

Table 3.12 Average IW_opt and EW scores: statistical accuracy (SA), informativeness for calibration questions (Inf(cal)) and all questions (Inf(all)) and combined score (Comb), by subregion. Last column depicts the percentage gain, with respect to the combined score, of the IW_opt over EW.

Ave scores by Region	IW_opt SA	IW_opt Inf(all)	IW_opt Inf(cal)	IW_opt Comb	EW SA	EW Inf(all)	EW Inf(cal)	EW Comb.	IW_opt Comb %gain over EW
AFR AB	0.399	2.691	0.821	0.301	0.458	1.069	0.336	0.155	94.7
AFR C	0.339	2.781	0.958	0.304	0.444	1.197	0.370	0.177	72.1
AFR D	0.243	2.579	0.818	0.175	0.396	1.050	0.346	0.138	26.4
AMR A	0.428	2.616	1.060	0.450	0.368	1.382	0.616	0.237	89.7
AMR B	0.382	2.290	0.829	0.307	0.388	1.248	0.481	0.188	63.7
AMR C	0.378	2.035	0.743	0.267	0.372	1.177	0.460	0.165	61.9
SEAR B	0.327	2.190	0.752	0.212	0.446	1.263	0.365	0.152	38.8
SEAR CD	0.277	2.193	0.831	0.188	0.453	1.046	0.404	0.178	5.5
EUR A	0.363	2.358	0.760	0.285	0.423	1.238	0.397	0.168	69.5
EUR B	0.276	2.242	0.702	0.168	0.407	1.013	0.319	0.119	41.1
EUR C	0.222	2.268	0.720	0.136	0.393	1.070	0.322	0.115	18.1
EMR A	0.410	2.887	0.834	0.342	0.463	1.349	0.337	0.156	119.5
EMR BC	0.319	2.574	0.908	0.278	0.426	1.235	0.371	0.159	74.8
EMR D	0.247	2.664	0.817	0.157	0.454	1.322	0.342	0.148	6.1
WPR A	0.464	2.272	0.767	0.357	0.447	1.236	0.340	0.137	160.2
WPR B	0.397	2.509	0.814	0.311	0.310	1.272	0.382	0.113	175.4
WPR C	0.396	2.589	0.898	0.311	0.313	1.239	0.373	0.107	191.9

In this Region averaging summary, IW_opt Comb. score is greater than that for EW for all the different Regions. In particular, the IW_opt Comb. score is more than doubled for EMR A (East Mediterranean A) and for the trio of West Pacific panels (WPR A, B and C); reasons for this performance in IW_opt DM pooling await further analysis.

Table 3.13. SA and Inf(all) scores averaged over all experts by hazard, with EW and IW_opt scores: : statistical accuracy (SA) and informativeness for all questions (Inf(all)). Scores for SARCO are highlighted in red to signal low mean statistical accuracy. The abbreviations are explained in Table 1 in Appendix A.

Hazard	Experts		DMs				
	SA	Inf(all)	EW SA	EW Inf(all)	IW_opt SA	IW_opt Inf(all)	
CAMP	0.037	2.432	0.514	0.921	0.474	2.122	*Diarrhoeal Enteric Hazards*
CRYP	0.057	2.873	0.546	1.197	0.62	3.03	
CYCL	0.006	4.017	0.617	2.07	0.365	4.623	
EAEC	0.027	2.835	0.506	1.102	0.451	2.528	
ENTA	0.009	3.352	0.577	1.855	0.628	4.102	
EPEC	0.014	2.281	0.488	0.753	0.335	2.244	
ETEC	0.033	2.358	0.468	0.824	0.298	2.094	
GIAR	0.057	3.006	0.549	1.368	0.517	3.441	
NORO	0.033	2.39	0.279	0.955	0.445	2.208	
ROTA	0.015	2.485	0.401	1.07	0.375	1.678	
NTS	0.055	1.785	0.498	0.567	0.456	1.138	
SHIG	0.05	2.178	0.458	0.84	0.429	1.585	

STEC	0.046	2.428	0.517	0.945	0.434	1.939	
VIBR	0.036	2.095	0.571	0.842	0.37	1.683	
BURC	0.017	3.142	0.457	1.407	0.384	2.665	*Non-Diarrhoeal Enteric Hazards*
Cbot	0.05	2.109	0.506	0.917	0.428	1.836	
CPERF	0.043	2.402	0.389	0.389	0.389	0.389	
Egan	0.034	1.64	0.349	1.083	0.347	1.438	
Emult	0.039	1.46	0.342	1.038	0.437	1.252	
HAV	0.034	2.572	0.569	1.117	0.338	2.595	
HEV	0.034	2.13	0.454	1.264	0.424	1.959	
LIST	0.049	2.535	0.494	0.888	0.535	2.075	
zMTBC	0.083	2.117	0.631	0.73	0.494	1.102	
PARA	0.002	3.042	0.352	1.162	0.28	2.952	
TYPH	0.004	3.363	0.353	1.277	0.363	3.047	
STAPH	0.026	3.208	0.337	1.17	0.37	2.089	
TOXO	0.037	2.378	0.541	1.086	0.199	2.081	
ANGIO	0.001	2.216	0.128	1.35	0.062	1.924	*Parasitic Hazards*
ASC	0.016	1.978	0.443	1.216	0.12	1.722	
FASC	0.032	2.557	0.35	1.392	0.215	2.861	
SARCO	**3.60E-06**	**1.407**	**0.002**	**0.966**	**9.80E-06**	**1.241**	
Tcru	0.011	3.51	0.494	1.853	0.304	2.893	
TOXOC	0.004	2.923	0.202	1.539	0.326	2.671	
TRICH	0.027	1.591	0.331	1.037	0.224	1.257	
AFB1	0.001	4.859	0.214	2.735	0.175	5.086	*Chemical Hazards*
iAs	0.005	2.825	0.194	1.024	0.167	2.743	
Cd	0.004	3.261	0.159	1.356	0.158	3.211	
DIOX	0.003	2.859	0.189	1.137	0.202	3.33	
Pb	0.017	2.234	0.356	0.687	0.207	1.793	
MeHg	0.004	3.445	0.291	1.9	0.162	3.239	
ave.	**0.026**	**2.607**	**0.403**	**1.176**	**0.337**	**2.347**	
stdev	**0.02**	**0.7**	**0.148**	**0.437**	**0.144**	**0.966**	
Pearson corr.	**-0.433**		**-0.141**		**-0.017**		
Rank corr.	**-0.401**		**-0.172**		**-0.118**		

In Table 3.13, the mean expert scores are the scores an analyst could expect when drawing a random expert for each hazard. The expected expert information score (2.607) is more than twice that of EW (1.176), meaning that the expected expert 90% uncertainty bands are roughly half the size of the EW 90% bands. The performance weighted combination IW_opt achieves informativeness (2.347) which is 90% of that expected from a random expert. The 90% uncertainty bands in this case are comparable to those of a random expert.

The oft-heard criticism -- that expert judgment studies can produce bloated confidence bands -- has some substance ***when applied to equally weighted combinations (EW) of expert uncertainties.*** However, this is NOT the case when applied to performance weighted combinations (PW). Indeed, the principal argument in favour of expert scoring and performance weighting is that this establishes confidence bands that are optimally constrained, coherent and logically defensible.is

An important feature is highlighted in the case SARCO. Here the mean SA scores for experts, EW and IW_opt are all disappointingly low; each of these combinations was unable to generate acceptable statistical performance. The objective of every SEJ application is to show that a structured statistical

combination of experts' judgments is not just a collection of ‘opinions', but a rational pooling, validated with calibration variables. This goal is meaningful only against the possibility that validation fails. If performance is not measured, then of course there can be no validation and hence no validation failures. Validation failure underscores the authenticity of validation successes.

However, while one hopes that validation failures are infrequent, an occasional failure need not entail disillusionment: in this instance, the problem owner should be advised that results for SARCO are not statistically reliable, and the protocol framing should be scrutinised.

Table 3.14 SA and Inf(all) scores averaged over all experts by subregion, with EW and IW_opt scores: statistical accuracy (SA) and informativeness for all questions (Inf(all)). Pearson and Spearman (rank) correlations are computed between average SA and information score for all questions, at the expert and DM level.

	Experts		**DMs**			
Region	**Ave SA**	**Ave Inf(all)**	**EW SA**	**EW Inf(all)**	**IW_opt SA**	**IW_opt Inf(all)**
AFR AB	0.023	2.477	0.458	1.069	0.399	2.691
AFR C	0.002	2.946	0.444	1.197	0.339	2.781
AFR D	0.024	2.12	0.396	1.05	0.243	2.579
AMR A	0.014	2.886	0.368	1.382	0.428	2.616
AMR B	0.011	2.468	0.385	1.256	0.373	2.326
AMR C	0.013	2.256	0.372	1.177	0.378	2.035
SEAR B	0.04	2.012	0.446	1.263	0.327	2.19
SEAR CD	0.038	2.327	0.453	1.046	0.277	2.193
EUR A	0.012	2.367	0.423	1.238	0.363	2.358
EUR B	0.037	1.901	0.407	1.013	0.276	2.242
EUR C	0.026	1.811	0.393	1.07	0.222	2.268
EMR A	0.021	2.409	0.463	1.349	0.41	2.887
EMR BC	0.011	2.427	0.426	1.235	0.319	2.574
EMR D	0.016	2.243	0.454	1.322	0.247	2.664
WPR A	0.004	2.211	0.447	1.236	0.464	2.272
WPR B	0.041	2.119	0.31	1.272	0.397	2.509
WPR C	0.044	1.951	0.313	1.239	0.396	2.589
Ave	0.022	2.29	0.409	1.201	0.345	2.457
Standard deviation	0.014	0.31	0.048	0.113	0.071	0.24
Pearson correlation	-0.647		-0.108		0.164	
Rank correlation	-0.646		-0.093		0.279	

Table 3.14 is similar to Table 3.13 but applied to regions instead of hazards. The aggregation by region is less granular than by hazard, and the poor performance of SARCO is not repeated. One interesting feature of Table 3.13 is accentuated in Table 3.14: the product moment (Pearson) and rank (Spearman) correlations between SA and Inf(all), while negative for experts, are attenuated for EW and become positive for IW_opt, though not statistically significant at the 5% level.

Together, Tables 3.11 through 3.14 highlight the value of performance weighting. The results, save SARCO, are validated and deliver informative and statistically accurate assessments.

3.4. Effects of training

The 2026 WHO study made a significant effort to provide experts with training in uncertainty quantification. The training materials have been detailed in Section 2.3.

The majority (72%) of the experts, involved in the 2026 elicitation participated in some form of training on this basis. Some only watched the training videos (V), while others watched the videos and answered the training questions, giving feedback on their assessments (TV). Moreover, certain experts accessed the training material multiple times.

From Table 3.11 we see that, overall, the training on offer improved the SA scores relative to WHO 2010 by an order of magnitude, with only modest effect on information scores. Since a substantial majority of experts participated in some form of training, score averages for the whole cadre of experts in 2026 do not differ much from the scores averaged over the experts who availed themselves of training materials.

Table 3.15 Effect of Training on DM scores. Statistical accuracy (SA) and informativeness for the calibration questions (Inf(cal)) are averaged (Ave) for experts who only watched training videos (V), who watched training videos and answered training questions (TV), who accessed the training materials multiple times (Multi Tr). The average scores for all the experts in the study (Ave all SA and Ave all Inf(cal)) and the average scores from the 2015 study (Ave SA 2015 and Ave Inf(cal) 2015). Number of expert assessments who covered different training materials.

Ave SA TV	**Ave SA V**	**Ave SA Multi Tr**	**Ave all SA**	**Ave SA 2015**
0.042	0.015	0.030	0.029	0.0032
Ave Inf(cal) TV	**Ave Inf(cal) V**	**Multi Tr**	**Ave all Inf(cal)**	**Ave Inf(cal) 2015**
1.259	1.380	1.313	1.317	1.5910
nr expert assessments				
1792	844	876	3215	
nr expts w ass'ts	**nr ass'ts**			
219	3215			

Feedback from experts regarding the training was very positive. These results encourage an expanded effort in expert training to enhance future online elicitations.

3.5. Source attribution results

After the collection of expert data, preliminary results were shared with the three taskforces. This was the case only for 29 of the 40 hazards, for which the source attribution results would be contributing to subsequent burden of disease calculations. The list of the 29 hazards is included in Pires et al. (2026).

The feedback from the taskforces included directly approving the results or providing an indication that certain point estimates were not in line with empirical data or otherwise seemed too low/high. As a consequence, follow-up responses were sought from some experts, who were subsequently contacted by the study team.

Furthermore, non-negligible assessments for blocked pathways/food groups were not accepted by the parasitic taskforce, and relevant experts were informed about this. All experts were asked if they wished to update their estimates, and some chose to do so, whereas some experts did not update their estimates and motivated their decision. The taskforces were presented with experts' feedback and, where

applicable, with their updated estimates. The compiled anonymized rationales (detailed per hazard and subregion) are included in Appendix F.

A notable example where experts argued for their choices and did not update their estimates is *Mycobacterium bovis/caprae/orygis*, which had been assumed to be 100% foodborne (as for the previous SEJ study). These experts' rationales were accepted by the taskforce, as well as the resulting estimates. Another notable example is Listeria, for which the same assumption held, i.e., 100% foodborne. For this case, experts' counter-rationales were nevertheless not accepted by the taskforce.

For the parasitic hazards, updated results were accepted by the taskforce only if assessments aligned with blocked pathways/food groups designations. A notable exception were results for *Toxoplasma gondii* and *Trichinella* spp. Finally, other outlier assessments were identified by the three taskforces, which were removed to obtain the approved results. More details are included in Pires et al. (2026).

In this manuscript, we do not make any evaluation of experts' professional knowledge. We report results obtained from aggregating all experts' assessments. When applicable, we use updated expert estimates. We note that some experts did not respond to the feedback request; however, their assessments are still reported and used. We document and indicate when experts' assessments are not in line with the taskforce designated blocked pathways/food groups (see Tables A5.1-A5.5 in Appendix A).

Experts' anonymized assessments can be found [include Github link here]. Source attribution results can be found [include Github link here].

3.6 Qualitative insights from 2015 to 2026

The 2015 attribution estimates updated to the current subregional classification was used in Pires et al. (2026) to compare the 2015 with the 2026 estimates. Moreover, while some experts took part in both studies, the majority of the current cadre of experts did not participate in the previous study.

4. Conclusions and discussion

This global study constitutes a significant methodological advance in the application of structured expert judgment. It is the largest study performed to date with 160 experts spread over 505 panels and a total of 3237 individual expert performance scores. The elicitation resulted in validated uncertainty quantification for source attribution for 40 hazards in 17 regions, roughly doubling the size of the previous 2015 study (Table 3.1). This up-scaling presented several challenges, the most important of which were: (a) the need to develop online training modules; (b) the need to use automated assessment tools, and (c) the need to develop dedicated processing software to handle this highly articulated data. In addition, the numbers of experts sought and required for each panel placed a significant burden on expert identification efforts.

Several steps were undertaken to address these challenges and ensure a rigorous process for such a large scale study. Intensive attention was given to the training steps, both for elicitors and experts. Elicitor training and preparation materials such as the elicitation script were essential to ensure consistent and effective elicitations. Expert training on the elicitation process and methodology was also extremely helpful and is seen to be associated with better performance in uncertainty quantification. For each broad hazard category, selected calibration questions were specific to each region. In the 2015 study, the calibration variables for the same hazard categories were divided into two regions only: developed and developing. This ensured a validating step tailored specifically to each panel. However, experts with knowledge of several regions carried a higher assessment burden, undertaken the calibration questions that were specific to each region separately.

The online elicitation tool developed in Qualtrics tailored questions to experts' regional and domain expertise, without burdening the preparation steps. Using Qualtrics however, might have been rather

difficult for experts when, for example, assessing multiple subregions. In this case, experts needed to navigate through the survey pages back and forth.

WHO intended to identify domain experts who were not engaged in the previous study. As a result, regionally diverse experts contributed to the study. This was different from typical studies, which generally adopt a more controlled approach to expert selection. Moreover, when expert coverage gaps were identified, targeted expert identification was engaged.

During the elicitation, experts self-selected the hazards/subregions they wished to assess, rather than having a specified list of hazards/subregions presented to them. Many experts showed considerable care when selecting the hazards/subregions and some even declined to provide assessments for some hazards initially chosen. A few experts did not provide any source attribution estimates despite completing the calibration questions.

Despite massive efforts to ensure subregional representativeness for each hazard, panel sizes remained, for many hazards/subregions, modest. Imputation methods were proposed and adopted to address expert coverage gaps or shortfalls, and regional, economic and global level proxies were obtained for hazards/subregions with insufficient assessments. It is noteworthy that the experts available for many panels, including merged assessments required for imputation, were still at the low end of what is generally considered viable in terms of numbers participating in the panel elicitations.

Despite the challenges mentioned above, the Classical Model was able to deliver validated uncertainty quantification, whose performance was superior to that of equal weighting (Tables 3.7, 3.8) and greatly superior to that of the 'average expert' (Tables 3.9, 3.10). Inherent in the notion of validation is that validation ***can*** fail. This is illustrated for the hazard Sarcocystis in Table 3.9: In this one case all experts produced statistically inaccurate assessments of calibration variables, and both the equal and the performance-based weighting were unable to achieve good statistical accuracy. This negative result illustrates the fact that validation is not a foregone conclusion. The gain in performance of performance weighting (IW_opt) over equal weighting (EW) is attested overall (Table 3.5), per hazard (Table 3.7) and per region (Table 3.8). The functioning of performance weighting is degraded by smaller panel sizes (Table 3.6).

Comparison between this study and the previous 2015 study presents a unique opportunity to assess the effects of expert training in uncertainty quantification. Experts who underwent training boosted their statistical accuracy scores to 0.03, compared to the average of 0.003 in 2015. Training lowered the average information score modestly, from 1.6 (2015) to 1.3 (2026). Of course, high information is a virtue only if combined with good statistical accuracy.

A non-prescriptive attitude was adopted for the way the experts approached source attribution assessments. While some pathways and food groups had been pre-identified as biologically implausible by the task forces, experts were informed that they could still assess those pathways/food groups, if they personally judged differently. This approach was embraced also in Sapp et al. (2022). Thus, some experts' assessments were not in agreement with some 'blocked' pathways, and the examples of *Mycobacterium bovis/caprae/orygis* and *Listeria monocytogenes*, mentioned in Section 2.5, are notable. We also note that for parasites, a stricter approach to blocking was applied by the task force, given the specifics of parasite life cycles. Nonetheless, we report and include any assessments of 'blocked' pathways/food groups as well, and noted where there are discrepancies with taskforce positions.

5. Acknowledgements

We would like to thank the WHO and FERG colleagues for their valuable support in identifying and approaching regional experts.

We would like to thank TU Delft research assistants who supported with the design of the study, elicitation and data collection: Femke Schürmann, Sofia Marques da Rocha Feliciano Pereira, Bodille Blomaard, Vangelis Nakos, Tyren Koning, Alessandra Primavera, Floor Jacobs, Si-Jing Chen, Judith Capel, Elina Ortolo, Natalia Vázquez Purriños, Elina Ortolo.

We would like to thank the engaged elicitors who supported with elicitations: Eduard Grau-Noguer, Zoe Baldwin, Stanley Chen, Uswatun Hasanah, Miranda Nonikashvili, Emi Grace Mary Gowshika, Maria Olorunsola, Sara Faife, Janet Rymound, Ankar Aggarwal, Pankaj Dhaka, Lisa O'Connor, Eiki Yamasaki, Alessandra Primavera, Muhammad Tanveer Munir, Kossi Brice Boris Legba, Abiodun Folake Abiola Omogoye, Yibaina Wang, Nada Alasiri, Reha Onur Azizoglu, Stephanie Poling, Selam Alemu, Ana Margarida Pignateli Vasconcelos de Assunção Alho, Dikshit Poudel, Dhanalakshmi Marimuthu, Jamila Seaton, Devin LaPolt, Sarah Hagan, Justine Alinaitwe, Emrecan Özeler, Belisário Moiane, Maria Francesca Iulietto, Malak Elbassuny.

6. References

Aspinall, W. P. & Cooke, R. Expert Elicitation and Judgement. in *Risk and Uncertainty assessment in Natural Hazards* (eds. J. C. Rougier, R. S. J. Sparks & L. J. Hill) 234–274 (Cambridge University Press, 2013).

Aspinall, W. P., Cooke, R. M., Havelaar, A. H., Hoffmann, S., & Hald, T. Evaluation of a Performance-Based Expert Elicitation: WHO Global Attribution of Foodborne Diseases. *PLoS One*, **11**(3), e149817 (2016).

Bamber J. L., Aspinall W. P. An expert judgement assessment of future sea level rise from the ice sheets. Nature Climate Change. 2013; 3: 424–427. doi: 10.1038/NCLIMATE1778

Baxter P. J., Aspinall W. P., Neri A., Zuccaro G., Spence R. J. S., Cioni R., Woo G. Emergency planning and mitigation at Vesuvius: A new evidence-based approach. Journal of volcanology and geothermal research. 2008; 178(3): 454–473. doi: 10.1016/j.jvolgeores.2008.08.015

Beshearse E., Bruce B. B., Nane G. F., Cooke R. M, Aspinall W., Hald T., et al. Attribution of Illnesses Transmitted by Food and Water to Comprehensive Transmission Pathways Using Structured Expert Judgment, United States. Emerg Infect Dis. 2021;**27**(1):182-195. https://dx.doi.org/10.3201/eid2701.200316

Butler, A. J., Thomas, M. K. & Pintar, K. D. M. Systematic Review of Expert Elicitation Methods as a Tool for Source Attribution of Enteric Illness. *https://home.liebertpub.com/fpd* **12**, 367–382 (2015).

Colonna, K. J., Nane, G. F., Choma, E. F., Cooke, R. M., & Evans, J. S. (2022). A retrospective assessment of COVID-19 model performance in the USA. *Royal Society open science*, *9*(10).

Cooke, R. (1991). *Experts in uncertainty: opinion and subjective probability in science*. Oxford university press.

Cooke, R. M., & Goossens, L. J. H. (1999). Procedures guide for structured expert judgment. *Project report to the European Commission, EUR*, *18820*.

Cooke, R. M., & Goossens, L. L. (2008). TU Delft expert judgment data base. *Reliability Engineering & System Safety*, *93*(5), 657-674.

Cooke, R. M. Goossens, L. J. H. (2000) Procedures guide for structured expert judgment. Project report EUR 18820EN, Nuclear Science and Technology, Specific Programme Nuclear fission safety 1994-98; Report to: European Commission. Luxembourg, Euratom. Also in Radiation Protection Dosimetry Vol. 90 No. 3.2000, 64 7, pp 303-311. See p. 30

Cooke, R. M., Marti, D. and Mazzuchi, T. A., (2021) Expert Forecasting with and without Uncertainty Quantification and Weighting: What Do the Data Say? International Journal of Forecasting, published online July 25, 2020, Fig. A.1Crotta, M. *et al.* Microbiological risk ranking of foodborne pathogens and food products in scarce-data settings. *Food Control* **141**, 109152 (2022).

Davydova, A. *et al.* Source attribution studies of foodborne pathogens, 2010–2023: a review and collection of estimates. *Food Microbiol.* **131**, 104812 (2025).

EFSA BIOHAZ Panel (EFSA Panel on Biological Hazards), 2013. Scientific Opinion on the evaluation of molecular typing methods for major food-borne microbiological hazards and their use for attribution modelling, outbreak investigation and scanning surveillance: Part 1 (evaluation of methods and applications). EFSA Journal 2013;11(12):3502,84 pp. doi:10.2903/j.efsa.2013.3502

Hald, T. *et al.* World Health Organization Estimates of the Relative Contributions of Food to the Burden of Disease Due to Selected Foodborne Hazards: A Structured Expert Elicitation. *PLoS One* **11**, e0145839 (2016).

Hanea, A. M., Nane, G. F., Bedford, T., & French, S. (Eds.). (2021). *Expert judgement in risk and decision analysis*. Cham, Switzerland: Springer.

Hanea, A. M., & Nane, G. F. (2021). An in-depth perspective on the classical model. In *Expert Judgement in Risk and Decision Analysis* (pp. 225-256). Cham: Springer International Publishing.

Havelaar, A. H. *et al.* (2015). World Health Organization Global Estimates and Regional Comparisons of the Burden of Foodborne Disease in 2010. *PLoS Med.* **12**, 1–23 (2015).

Hoffmann, S. *et al.* Attribution of global foodborne disease to specific foods: Findings from a World Health Organization structured expert elicitation. *PLoS One* **12**, e0183641 (2017).

Mughini-Gras, L. *et al.* Critical Orientation in the Jungle of Currently Available Methods and Types of Data for Source Attribution of Foodborne Diseases. *Front. Microbiol.* **10**, 475117 (2019).

Pires, S. M. *et al.* (2009). Attributing the human disease burden of foodborne infections to specific sources. *Foodborne Pathog. Dis.* **6**, 417–424 (2009).

Pires, S. M. *et al.* (2026). World Health Organization Attribution of Burden of Foodborne Diseases to Transmission Pathways and Specific Foods. *Submitted for publication.*

Qualtrics. Qualtrics XM Platform. Preprint at (2023).

R Core Team. R Foundation for Statistical Computing. Preprint at (2021).

Sapp, A. C., Amaya, M. P., Havelaar, A. H., & Nane, G. F. (2022). Attribution of country level foodborne disease to food group and food types in three African countries: Conclusions from a structured expert judgment study. PLoS neglected tropical diseases, 16(9), 1-21. Article e0010663. https://doi.org/10.1371/JOURNAL.PNTD.0010663

WHO estimates of the global burden of foodborne diseases. *WHO* (2015).

WHO Global Strategy for Food Safety 2022-2030 : Towards Stronger Food Safety Systems and Global Cooperation. (World Health Organization, Geneva, 2022)

WHO. Call for experts on source attribution of foodborne disease hazards. https://www.who.int/news-room/articles-detail/call-for-experts-on-source-attribution-of-foodborne-disease-hazards (2023)

Appendices

Appendix A Background documents

Table A1. List of the 40 elicited hazards and abbreviations.

Enteric diseases (diarrheal diseases)	**Enteric diseases (non-diarrheal diseases)**	**Parasitic diseases**	**Chemicals and Toxins**
Campylobacter spp. **(CAMP)**	*Brucella* spp. **(BRUC)**	*Ascaris lumbricoides* **(ASC)**	Aflatoxin B1 **(AFB1)**
Cryptosporidium spp. **(CRYP)**	*Clostridium botulinum* **(Cbot)**	*Echinococcus multilocularis* **(Emult)**	Dioxin & DL-PCBs **(DIOX)**
Cyclospora **(CYCL)**	*Clostridium perfringens* **(CPERF)**	*Echinococcus granulosus* **(Egan)**	Lead **(Pb)**
Entamoeba histolytica **(ENTA)**	Hepatitis A virus **(HAV)**	*Trypanosoma cruzi* **(Tcru)**	Methyl Mercury **(MeHg)**
Enteroaggregative *E.coli* **(EAEC)**	Hepatitis E virus **(HEV)**	*Fasciola* spp.& *Fasciolopsis* **(FASC)**	Cadmium **(Cd)**
Enteropathogenic *E.coli* **(EPEC)**	*Listeria monocytogenes* **(LIST)**	*Toxoplasma* gondii **(FASC)**	Arsenic **(iAs)**
Enterotoxigenic *E.coli* **(ETEC)**	*Mycobacterium bovis/caprae/orygis* **(zMTBC)**	*Trichinella spp*. **(TRICH)**	
Giardia spp. **(GIAR)**	*Salmonella* Paratyphi A **(PARA)**	Toxocara **(TOXOC)**	

Norovirus **(NORO)**	*Salmonella* Typhi **(TYPH)**	Angiostrongylus **(ANGIO)**	
Rotavirus **(ROTA)**	Bacterial toxins: Staph. Aureus **(STAPH)**	Sarcocystis **(SARCO)**	
Non-typhoidal *Salmonella enterica* **(NTS)**			
Shigella spp. (**SHIG)**			
Shiga toxin-producing *E.coli* **(STEC)**			
Vibrio cholerae **(VIBR)**			

Table A2. List of countries assigned to the 17 subregional clusters.

Region Code	Countries
AFR AB	Botswana, Equatorial Guinea, Gabon, Mauritius, Namibia, Seychelles, South Africa
AFR C	Algeria, Angola, Benin, Cameroon, Cabo Verde, Comoros, Congo, Côte d'Ivoire, Eswatini, Ghana, Guinea, Kenya, Lesotho, Mauritania, Nigeria, São Tomé and Príncipe, Senegal, United Republic of Tanzania, Zambia, Zimbabwe
AFR D	Burkina Faso, Burundi, Central African Republic, Chad, Democratic Republic of Congo, Eritrea, Ethiopia, Gambia, Guinea-Bissau, Liberia, Madagascar, Malawi, Mali, Mozambique, Niger, Rwanda, Sierra Leone, South Sudan, Togo, Uganda
AMR A	Antigua and Barbuda, Bahamas, Barbados, Canada, Chile, Guyana, Panama, Saint Kitts and Nevis, Trinidad and Tobago, United States of America, Uruguay
AMR B	Argentina, Belize, Brazil, Colombia, Costa Rica, Cuba, Dominica, Dominican Republic, Ecuador, El Salvador, Grenada, Guatemala, Jamaica, Mexico, Paraguay, Peru, Saint Lucia, Saint Vincent and the Grenadines, Suriname
AMR C	Bolivia (Plurinational State of), Haiti, Honduras, Nicaragua, Venezuela (Bolivarian Republic of)
SEAR B	Indonesia, Maldives, Thailand
SEAR CD	Bangladesh, Bhutan, Ina, Myanmar, Nepal, Democratic People's Republic of Korea, Sri Lanka, Timor-Leste
EUR A	Andorra, Austria, Belgium, Croatia, Cyprus, Czechia, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Latvia, Lithuania, Luxembourg, Malta, Monaco, Netherlands (Kingdom of the), Norway, Poland, Portugal, Romania, San Marino, Slovakia, Slovenia, Spain, Sweden, Switzerland, United Kingdom of Great Britain and Northern Ireland
EUR B	Albania, Armenia, Azerbaijan, Belarus, Bosnia and Herzegovina, Bulgaria, Georgia, Kazakhstan, Montenegro, North Macedonia, Republic of Moldova, Russian Federation, Serbia, Türkiye, Turkmenistan

Region Code	Countries
EUR C	Kyrgyzstan, Tajikistan, Ukraine, Uzbekistan
EMR A	Bahrain, Kuwait, Oman, Qatar, Saudi Arabia, United Arab Emirates
EMR BC	Djibouti, Egypt, Iran (Islamic Republic of), Iraq, Jordan, Lebanon, Libya, Morocco, Pakistan, Tunisia
EMR D	Afghanistan, Somalia, Sudan, Syrian Arab Republic, Yemen
WPR A	Australia, Brunei Darussalam, Cook Islands, Japan, Nauru, New Zealand, Niue, Singapore, Republic of Korea
WPR B	China, Fiji, Malaysia, Marshall Islands, Palau, Tonga, Tuvalu
WPR C	Cambodia, Kiribati, Lao People's Democratic Republic, Micronesia (Federated States of), Mongolia, Papua New Guinea, Philippines, Samoa, Solomon Islands, Vanuatu, Viet Nam

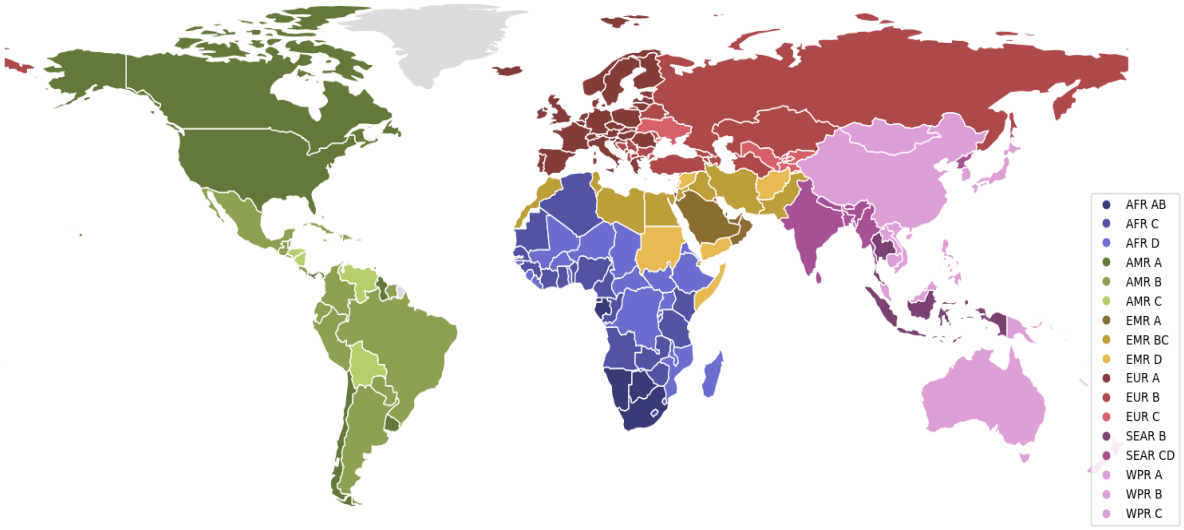

Figure A1. World map of the 17 subregions classification.

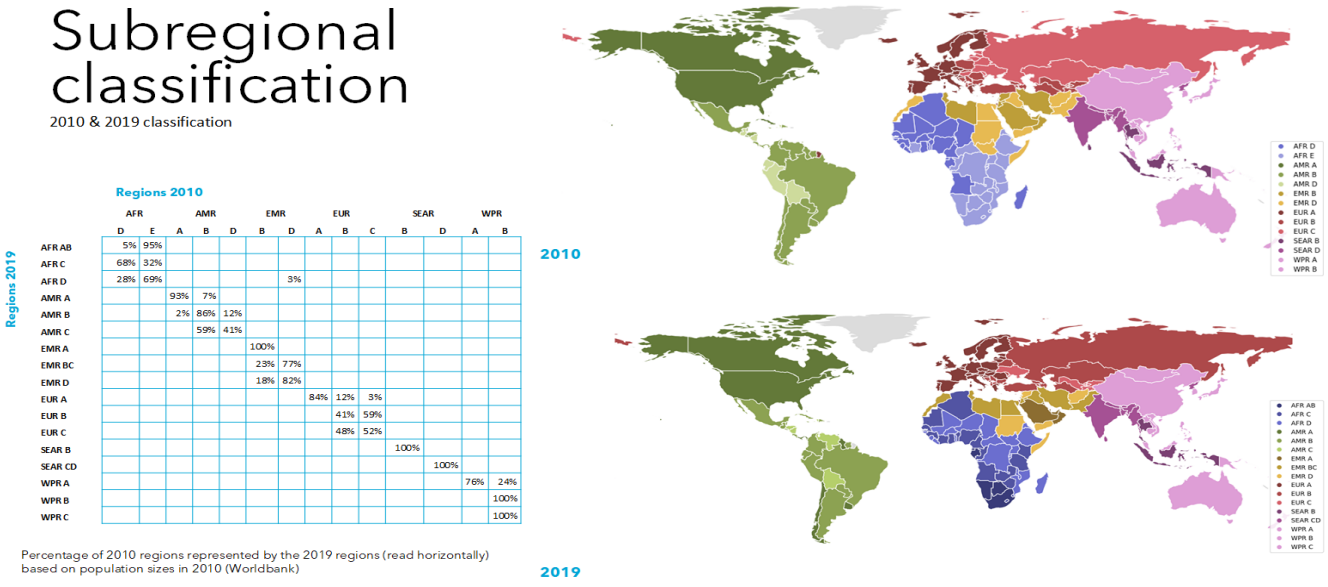

Figure A2. 2026 and 2015 world map classifications, including the correspondence between the 2026 and 2015 regions, expressed as percentage of 2015 subregions (with respect to population size, Worldbank) represented by the 2026 subregions.

Table A3. General factors of potential influencing factors to lead to substantial changes in hazard source attribution estimates.

General factors:	Improved sanitation • Specific (food safety/public health) policies • Epidemiological changes (changes in eating/dietary habits, lifestyle, risk factors, exposure patterns, etc.) • Environmental changes (changes in climate, weather, landscape, urbanization degree, etc.) • Farming / food processing systems, including market changes (such as import/export)
With more insight into:	Changes in food systems, e.g. - increased reliance on imports - changes in food markets - shifts away from home production - increased consumption of food eaten away from home - investments in the safety of open markets - changes in food marketing, for example, growth of supermarket chains - improvements in cold-chain
	Demographic shifts, e.g., urbanization, changes in the population's age profiles, etc. Changes in surveillance methods and data - Use of multiplex molecular methods means that more samples are analysed for more pathogens (i.e., meaning change in diagnostic algorithms rather than greater sensitivity, although that could also contribute). This has certainly impacted on greater realization of how much Crypto there is in Europe. Environmental changes, e.g., - warming waters and the rise of aquaculture - more flooding resulting in contamination of fresh produce and, conversely, water shortage resulting in use of potentially unsafe water for irrigation - greater encroachment on wilderness areas and fragmentation of wilderness

	- (changes in climate, weather, landscape, urbanization degree, etc.)
For chemical hazards:	• Regulatory improvement: updates in food safety regulations and standards may impact the use and control of foodborne chemicals. • Changes in monitoring methods and data: Technological advancements: advances in technology might have led to more precise methods of detecting and monitoring foodborne chemicals. • Advancements in lab tests and analytical Techniques. • Consumer awareness. • Changes in food systems: changes in the global food supply chain, with increased imports and exports, may alter the types and amounts of chemicals used in different regions • Environmental factors. • Economic Factors

Table A4. Transmission pathway definitions and food groups considered.

Transmission pathway	**Definition**
Food*	Consumption of contaminated food, including non-water beverages such as juices. Experts were instructed to consider contamination of food until the point where the food entered the place where it was prepared for consumption (e.g., kitchen). This excluded cross-contamination during food preparation.
Contact with Animals (Domestic, farmed, or wild)	Direct contact with an animal or its bodily fluids (excluding raw milk or other fluids consumed as food), fur, hair, feathers, scales, or skin, or by contact with the local environment where an infected animal, its visible excreta, fur, hair, feathers, scales, or skin was simultaneously present with the exposed person (e.g., barns, petting zoos, and pet stores)
Human-to-human contact	Direct contact with infected persons or their bodily fluids, by contact with the local environment where an exposed person is simultaneously present with an infected person, or through vertical transmission (i.e., from mother to child)
Water	Consumption of water, direct contact with water, or inhalation of aerosols originating from water. Includes drinking water, bottled water, recreational water (treated and untreated), and other water sources
Soil	Exposure to contaminated soil or mud

Other	Any other transmission not attributable to foodborne, waterborne, person-to-person, or animal contact transmission as defined above (e.g., airborne)
Specific foods	Beef, small ruminant meat, pig meat, poultry meat, game, dairy, eggs, vegetables, fruits and nuts, grains and beans, oils and sugar, finfish, shellfish, seaweed, consumed raw or cooked

Table A5.1. Blocked pathways enteric diarrhoeal. Part 1.

Transmission pathway	***Campylobacter* spp.**	***Cryptosporidium* spp.**	***Cyclospora***	***Entamoeba histolytica***	**Enteroaggregative *E.coli* (EAggEC)**	**Enteropathogenic *E.coli* (EPEC)**	**Enterotoxigenic *E.coli* (ETEC)**
Food	X	X	X	X	x	X	x
Contact with animals (domestic or wild)	X	X	X		X	X	X
Human contact	X	X	X	X			
Water	X	X	X	X	X	X	X
Soil	X	X	X	X	X	X	X
Other (e.g., airborne/pollution, occupational)	X	X	X	X	X	X	X
Specific foods							
Beef	X	X	x	x	X	X	X
Small ruminants' meat	X	X	X	X	X	X	X
Pigs' Meat	X	X	X	X	X	X	X
Poultry Meat	X	X	X	X	X	X	X
Game	X	X	X	X	X	X	X
Dairy	X	X	X	X	X	X	X
Eggs							
Vegetables	X	X	X	X	X	X	X
Fruits and Nuts	X	X	X	X	X	X	X
Grains and Beans	X				X	X	X
Oils and sugar	X				X	X	X
Finfish	X	X		X	X	X	X
Shellfish	X	X	X	X	X	X	X
Seaweed	X	X	X	X	X	X	X

Table A5.2. Blocked pathways enteric diarrhoeal. Part 2.

Transmission pathway	***Giardia* spp.**	**Norovirus**	**Rotavirus**	**Non-typhoidal *Salmonella enterica***	***Shigella* spp.**	**Shiga toxin-producing *E.coli* (STEC)**	***Vibrio cholerae***
Food	X	X	X	X	X	X	X
Contact with animals (domestic or wild)	X			X	X	X	X
Human contact	X	X	X	X		X	X
Water	X	X	X	X	X	X	X
Soil	X	X	X	X	X	X	X
Other (e.g., airborne/pollution, occupational)	X	X	X	X	X	X	X
Specific foods							
Beef	X	X	X	X	X	X	X
Small ruminants' meat	X	X	X	X	X	X	X
Pigs' Meat	X	X	X	X	X	X	X
Poultry Meat	X	X	X	X	X	X	X
Game	X	X	X	X	X	X	X
Dairy	X	X	X	X	X	X	X
Eggs				X	X		
Vegetables	X	X	X	X	X	X	X
Fruits and Nuts	X	X	X	X	X	X	X
Grains and Beans				X	X	X	
Oils and sugar				X	X	X	
Finfish	X	X	X	X	X	X	X
Shellfish	X	X	X	X	X	X	X
Seaweed	X	X	X	X	X	X	X

Table A5.3. Blocked pathways enteric non-diarrhoeal.

Transmission pathway	*Brucella* spp.	*Clostridium perfringens*	*Clostridium botulinum*	Hepatitis A virus	*Listeria monocyto genes*	*Mycobacterium bovis*	*Salmonella enterica* Paratyphi A	*Salmonella enterica* Typhi	Hepatitis E virus	Bacterial toxins: *Staph. Aureus*
Food	X	X	X	X	X	X	X	X	X	X
Contact with animals (domestic or wild)	X			X	X	X			X	
Human contact				X		X	x	X	X	
Water	X			X	X	X	X	X	X	
Soil	X			X	X	X	X	X		
Other (e.g., airborne/pollution, occupational)	X			X	X	X	X	X		
Specific foods										
Beef	X	X	X	X	X	X	X	X		X
Small ruminants' meat	X	X	X	X	X	X	X	X		X
Pigs' Meat	X	X	X	X	X	X	X	X	X	X
Poultry Meat		X	X	X	X	X	X	X		X
Game	X	X	X	X	X	X	X	X	x	X
Dairy	X	X	X	X	X	X	X	X		X
Eggs			X		X					
Vegetables			X	X	X	X	X	X		X
Fruits and Nuts			X	X	X	X	X	X		X
Grains and Beans			X		X	X				X
Oils and sugar					X	X				X
Finfish			X	X	X	X	X	X		X
Shellfish			X	X	X	X	X	X		
Seaweed			X	X	X	X	X	X		

Table A5.4. Blocked pathways parasites.

Transmission pathway	***Toxoplasma gondii***	***Fasciola & Fasciolopsis***	***Echinococcus Multilocularis***	***Echinococcus Granulosus***	***Trypanosoma cruzi***	***Trichinella* spp.**	***Ascaris lumbricoides***	***Toxocara***	***Angiostrongylus***	***Sarcocystis***
Food	X	X	X	X	X	X	X	X	X	X
Contact with animals (domestic or wild)	X		X	X	X					
Human contact					X					
Water		X	X	X						X
Soil	X	X	X	X				X		X
Vector borne					X		X			
Other (e.g. airborne/pollution/ occupational)					X		X	X		
Specific foods										
Beef					X			X		X
Small ruminants' meat					X			X		
Pigs' Meat					X	X		X		X
Poultry Meat					X			X		
Game					X	X	X			
Dairy					X					
Eggs										
Vegetables	X	X	X	X	X		X	X	X	
Fruits and Nuts	X	X	X	X	X		X	X	X	
Grains and Beans										
Oils and sugar										
Finfish										
Shellfish (including snails)					X				X	
Seaweed										

Table A5.5. Blocked pathways chemicals.

Major pathways	Aflatoxin B1	Dioxin & DL-PCBs	Lead	Methyl Mercury	Cadmium	Arsenic
Food	X	X	X	X	X	X
Contact with animals (domestic or wild)			X			
Human contact			X			
Water			X	X		X
Soil		X	X	X		
Other (e.g. airborne/pollution/ occupational)		X	X	X	X	
Specific foods						
Beef		X	X		X	X
Small ruminants' meat		X	X		X	X
Pigs' Meat		X	X		X	X
Poultry Meat		X	X		X	X
Game		X	X		X	X
Dairy		X	X		X	X
Eggs		X	X		X	X
Vegetables		X	X		X	X
Fruits and Nuts	X	X	X		X	X
Grains and Beans	X	X	X		X	X
Oils and sugar		X	X		X	X
Finfish		X	X	X	X	X
Shellfish		X	X	X	X	X
Seaweed		X	X		X	X

Appendix B Elicitation materials

Table B1.1. Calibration questions overview for enterics, where, for each hazard, 13 calibration questions have been selected.

	1. Food Supply					2. Health and diarrheal disease			3. Outbreak and disease surveillance																		
									3.1 Europe				3.2 America				3.3 Asia		3.4 Oceania				3.5 Africa			3.6 Mi
Sub-Regions	**1.1**	**1.2**	**1.3**	**1.4**	**1.5**	**2.1**	**2.2**	**2.3**	**3.1.1**	**3.1.2**	**3.1.3**	**3.1.4**	**3.2.1**	**3.2.2**	**3.2.3**	**3.2.4**	**3.3.1**	**3.3.2**	**3.4.1**	**3.4.2**	**3.4.3**	**3.4.4**	**3.5.1**	**3.5.2**	**3.5.3**	**3.6.1**
AFR AB	x		x	x	x	x	x	x															x	x		
AFR C	x		x	x	x	x	x	x															x	x		
AFR D	x		x	x	x	x	x	x															x	x		
AMR A	x		x	x	x			x					x	x	x	x										
AMR B	x		x	x	x	x	x	x					x			x										
AMR C	x		x	x	x	x	x	x					x			x										
EMR A	x		x	x	x	x	x	x																		x
EMR BC	x		x	x	x	x	x	x																		x
EMR D	x		x	x	x	x	x	x																		x
EUR A	x		x	x	x			x	x	x	x	x														
EUR B	x		x	x	x			x	x	x	x	x														
EUR C	x		x	x	x			x	x	x	x	x														
SEAR B	x		x	x	x	x	x	x									x	x								
SEAR CD	x		x	x	x	x	x	x									x	x								
WPR A	x		x	x	x			x											x	x	x	x				
WPR B	x		x	x	x	x	x	x											x	-	x	-				
WPR C	x		x	x	x	x	x	x											x	-	x	-				

Table B1.2. Calibration questions overview for parasites, where, for each hazard, 13 calibration questions have been selected.

	1. Food Supply					2. Health and diarrheal disease		3. Outbreak and disease surveillance																				4.
								3.1 Europe				3.2 America				3.3 Asia		3.4 Oceania			3.5 Africa		3.6 Middle East					
	1.1	1.2	1.3	1.4	1.5	2.1	2.2	2.3	3.1.1	3.1.2	3.1.3	3.1.4	3.2.1	3.2.2	3.2.3	3.2.4	3.3.1	3.3.2	3.4.1	3.4.2	3.4.3	3.4.4	3.5.1	3.5.2	3.5.3	3.6.1	3.6.2
	x		x	x	x			x		x			x		x										x		
ciolopsis	x		x	x	x			x		x			x		x										x		
spp	x		x	x	x			x		x			x		x										x		
	x		x	x	x			x		x			x		x										x		

cruzi	x		x	x	x			x		x			x		x										x		
	x		x	x	x			x		x			x		x										x		
us	x		x	x	x			x		x			x		x										x		
coides	x		x	x	x			x		x			x		x										x		
um	x		x	x	x			x		x			x		x										x		
	x		x	x	x			x		x			x		x										x		
	x		x	x	x			x		x			x		x										x		
	x		x	x	x			x		x			x		x										x		
	x		x	x	x			x		x			x		x										x		

Table B1.3. Calibration questions overview for chemicals, where, for each hazard, 13 calibration questions have been selected.

	1. Food Supply					2. Health and diarrheal disease			5. Chemicals											
Chemical	**1.1**	**1.2**	**1.3**	**1.4**	**1.6**	**2.1**	**2.2**	**2.3**	**5.1**	**5.2**	**5.3**	**5.4**	**5.5**	**5.6**	**5.7**	**5.8**	**5.9**	**5.10**	**5.11**	**5.12**	**5.13**
Aflatoxin B1	x	x	x	x	x			x				x		x	x	x					
Dioxin & DL-PCBs	x		x	x	x			x			x	x				x				x	x
Lead	x		x	x	x			x	x	x		x	x			x	x	x	x		
Methyl Mercury	x		x	x	x			x	x	x		x	x			x	x	x	x		
Cadmium	x		x	x	x			x	x	x		x	x			x	x	x	x		
Arsenic	x		x	x	x			x	x	x		x	x			x	x	x	x		

Table B2. List of all calibration questions.

Calibration questions for World Health Organization Source Attribution SEJ study

1. Questions about food supply

1.1. The FAOSTAT statistics provided by the Food and Agricultural Organization (FAO) allow calculation of the change in the percentage of a country's domestic food supply over time.

Across the region of South-eastern Asia*, the total amount of poultry meat supplied domestically in 2015 was 8,323 thousand tons. By 2019, what was the percentage change in weight of domestic poultry meat supply in this region relative to 2015?

Please, express your answer as a percentage point difference.

- Use this equation:
    - *((value 2019−value 2015)/value 2015)*100*

*Different regional variants of this question were available to the experts

1.2. The FAOSTAT statistics provided by the Food and Agricultural Organization allow calculation of the change in the percentage of a country's export quantity of specific food items over time.

Across the African* regions, the total amount of maize and products thereof exported by all countries in 2015 was 2,476 thousand tonnes. What was the percentage change in this total weight from 2015 to 2019?

Please, express your answer as a percentage point difference.

- Use this equation:
    - *((value 2019−value 2015)/value 2015)*100*

*The different regional variants of this question are available at the end of this document.

1.3. The FAOSTAT statistics provided by the Food and Agricultural Organization allow for the comparison of annual meat consumption across countries. Meat in this context includes meat from bovines, sheep, goats, pigs, and poultry.

Expressed in kilo per capita per year (kg/capita/yr), what was the absolute difference between the country with lowest and highest consumption of meat (all types) in 2019 in the world?

Please, express your answer as a number in kg/capita/yr.

- Use this equation:
    - *Highest meat consumption (kg/capita/yr) – lowest meat consumption (kg/capita/yr)*

1.4. One available measure of the importance of animal products in a population's diet is the percentage contribution of animal products out of the total food supply available for human consumption measured as kcal per capita day.

FAOSTAT statistics provided by the Food and Agricultural Organization report the percentage of animal-based foods out of the total national food supply (*kcal per capita per day*) that were available for human consumption in 2019, by country. Think of the two countries in South America* with the highest and lowest such percentages. What was the difference in percentages between these two countries?

Please, express your answer as a positive percentage point difference.

- Use this equation:
    - *(highest % per capita value – lowest % per capita value)*

*The different regional variants of this question are available at the end of this document.

1.5. The FAOSTAT statistics provided by the Food and Agricultural Organization allow for the comparison of protein supply from fish and seafood measured as gram per capita per day (g/capita/day) across countries.

In the Oceanian region* in 2019, fish and seafood on average contributed 6.22 g protein per capita per day. Expressed in gram protein per capita per day (g/capita/day), what was the absolute difference between the country with highest and lowest intake of protein from fish and seafood in the Oceanian region in 2019?

Please, express your answer as a positive number in g/capita/day.

- Use this equation:
    - *Highest protein from seafood (g/capita/day) – lowest protein from seafood (g/capita/day)*

*The different regional variants of this question are available at the end of this document.

2. Questions about health and diarrheal disease

Background: Diarrhea is a leading cause of death of children, accounting for approximately 9 per cent of all deaths among children under five years of age worldwide in 2019. This translates to over 1,300 young children dying each day, or about 484,000 children a year.

The Institute of Health Metric and Evaluation (IHME) shares the data used for the Global Burden of Disease (GBD) study and thus publishes country-specific estimates of diarrheal deaths expressed as number of deaths per 100,000 inhabitants. These estimates indicate that many Low and Middle-Income Countries (LMIC) have experienced a decline in diarrheal deaths in the period from 2000 to 2019. For the total African region, the average decrease in the same period was 81.01 cases per 100,000 – being reduced from 148.13 cases per 100,000 in 2000 to 67.12 in 2019.

2.1. Rotavirus is an important cause of diarrheal disease in LMIC. Implementation of a Rotavirus vaccine in the national child immunization programs has led to a reduction in diarrheal incidence and mortality in many countries. According to the WHO, by the end of 2018, 101 countries were using the rotavirus vaccine. Of these, 75 LMIC reported data on the share of one-year old children that were vaccinated against Rotavirus in 2019.

Based on WHO and UNICEF estimates, how many of the 75 countries for which there is data had a vaccination coverage of less than 75% of one-year old children in 2019?

Please, express your answer as an integer *(# countries)*.

2.2. Oral rehydration therapy (ORT) is another intervention that has contributed to the reduction of diarrheal disease and mortality in LMIC. Since 2004, UNICEF and WHO have recommended treating childhood diarrhea with a combination of oral rehydration salts (ORS) and zinc as a proven and affordable treatment.

Based on data from UNICEF analysed in the period 2015-2021, the South Asia region had the highest coverage with 20.7% of children under five years of age being treated for diarrhea with ORS+zinc. What was the coverage percentage for children under 5 years of age in Sub-Saharan Africa (SSA)?

Please, express your answer in percentage.

- Use this equation:
    - *(# children < 5 yr with diarrhea treated in SSA / # children < 5 with diarrhea in SSA)*100*

2.3 According to data published by the World Bank, the prevalence of underweight children under 5 years of age was 16.3% in 2010. The prevalence of underweight children is here defined as the percentage of children under age 5 whose weight for age is more than two standard deviations below the median for the international reference population ages 0-59 months. The data are based on the WHO's 2006 Child Growth Standards.

Think of the country with the highest prevalence of underweight children in 2010. What was the percentage change in underweight prevalence from 2010 to 2019 in this country?

Please, express your answer in percentage change.

- Use this equation:
    - *((% underweight children < 5 yr in 2019 − % underweight children < 5 yr in 2010) / (% underweight children < 5 yr in 2010)*100*

3. Questions Based on Disease and Outbreak Surveillance Data

3.1 Questions based on EU disease surveillance reports

Background: All EU member states collect active surveillance data on several major zoonoses and on foodborne outbreaks (Zoonoses Directive 2003/99/EC). They report this information to the European Food Safety Authority (EFSA) annually. EFSA, in collaboration with the European Centre for Disease Prevention and Control (ECDC), publishes an annual summary of this data, usually released by the end of the following year. The following questions refer to data published in the EFSA reports.

3.1.1. The rate of reported outbreaks per 100,000 population decreased from levels observed before the COVID-19 pandemic. Think about the mean of the annual rate per 100,000 in the period 2017-19, and estimate the percentage decline in 2020, relative to the 2017-19 mean rate.

Please express your answer in positive percentage points.

- Use this equation:
    - *((mean of the annual rate per 100,000 2017–2019 – annual rate per 100,000 2020)/ mean of the annual rate per 100,000 2017–2019)*100*

3.1.2. According to the Zoonoses Directive 2003/99/EC, EU member states (MS) must report bovine brucellosis annual monitoring data. The reports submitted by the EU MS are based on Council Directive 64/432/EEC and subsequent legislation and are essential for the assessment of the epidemiological situation in MS and MS regions, when declared officially brucellosis free in cattle (OBF).

In 2019, 20 OBF and 8 non-OBF EU MS reported a total of 489 cattle herds infected with *Brucella spp*. How many of these herds were from non-OBF regions?

Please express your answer in positive integers *(# herds)*.

3.1.3. In 2019, 220,682 confirmed cases of human campylobacteriosis were reported by the 28 EU member states, corresponding to an EU mean notification rate of 59.7 cases per 100,000 population.

Think of the EU member state with the highest notification rate and indicate your estimate of the difference of this rate from that of the EU mean rate.

Please express your answer in numbers per 100,000 population.

- Use this equation:
    - *(highest incidence rate (# cases per 100,000) – EU mean (59.7 cases per 100,000))*

3.1.4. In 2019, 22 EU member states provided data from their national control program on *Salmonella* in broiler flocks. Commission Regulation (EU) No 200/2012 requires EU member states (MS) to report separately the results obtained by the Food Business Operator (FBO) and by the Competent Authority (CA).

Considering the data from the 18 EU MS that reported separate results from both CA and FBO, the combined prevalence of *Salmonella* enteritidis and *Salmonella* typhimurium, the target serovar, was 0.09% of 246,083 flocks tested by the FBO. What was the prevalence in percentage in the 5,013 broiler flocks tested by the CA?

Please express your answer in percentage.

- Use this equation:
    - *(# infected flocks / 5,013 flocks tested by the CA)*100*

3.2 Questions based on US disease surveillance data as reported by CDC, NNDSS and FoodNet

Background: Since 1996, the U.S. Foodborne Diseases Active Surveillance Network, or FoodNet, has been collecting and presenting data on infections commonly transmitted through food. FoodNet's surveillance area covers 51 million people, representing 15% of the U.S. population. Results are reported annually. In 2016, FoodNet launched FoodNet Fast, which is an online toolbox for accessing information reported to FoodNet.

3.2.1. From 2016 to 2019, there was an increase in the number of human cases of *Cyclospora* reported through the FoodNet sites. In 2016, 55 cases were reported corresponding to an incidence of 0.11 per 100,000 population. What was the incidence rate of human *Cyclospora* infections reported in 2019?

Please express your answer as cases per 100,000.

3.2.2. Hemolytic Uremic Syndrome (HUS) is a serious condition that can lead to kidney failure, permanent health problems, and even death. HUS is most often caused by Shiga toxin-producing Escherichia coli (STEC) infection, in particular by STEC O157. FoodNet conducts active surveillance for physician-diagnosed, pediatric HUS cases. In the 10-year period from 2012 to 2021, a total of 663 HUS cases in children under 18 years (0.6 per 100,000 children) were reported from FoodNet sites.

How many of these HUS infections were in children between 10 and 18 years of age?

Please express your answer as positive integers *(# cases)*.

3.2.3. In 2019, a total of 13,979 cases of human cryptosporidiosis were reported through the NNDSS by CDC. This corresponds to an incidence of 4.3 cases per 100,000. What was the incidence reported for the jurisdiction/state with the highest incidence?

Please express your answer as cases per 100,000.

Background: The U.S. "National Notifiable Disease Surveillance System" (NNDSS) tracks infectious diseases that laboratory professionals and doctors are required to report to the state or territorial public health agency. These agencies voluntarily submit the information to NNDSS, which the Centers for Disease Control and Prevention (CDC) oversees.

3.2.5. From 2016 to 2019, the incidence of Hepatitis A virus (HAV) infections increased from 0.6 to 5.7 cases per 100,000 as reported by the CDC. Infections among white non-Hispanic persons increased the most in this period. By how many folds did this incidence rate in white non-Hispanic increase from 2016 to 2019?

Please express your answer as a positive integer.

- Use this equation:
    - *(incidence rate 2019/incidence rate 2016)*

3.3 Questions based on disease surveillance data reported in Asia

3.3.1. Since 2010, the China National Center for Food Safety Risk Assessment has established the Foodborne Disease Outbreak Surveillance System (FDOSS) to monitor foodborne disease outbreaks.

From 2010 to 2020, the FDOSS gathered data from 18,331 outbreaks. What percentage of these outbreaks was associated with meat and meat products?

Please express your answer as a percentage.

- Use this equation
- *(# outbreaks associated with meat / # total outbreaks)*100*

3.3.2. The Integrated Disease Surveillance Programme (IDSP) in India publishes reported surveillance data in the IDSP monthly newsletter. According to the report publishing data from August 2019, 1,999,096 episodes of acute diarrheal disease (ADD) were reported in this month.

A total of 4,344 samples were tested for enteric pathogens. How many of these samples tested positive for cholera?

Please express your answer as an integer, number of samples.

3.4 Questions based on disease surveillance data as reported by New Zealand and Australia

Background: The following questions are based on data published in the "Annual reports concerning foodborne disease in New Zealand 2019". The reports are prepared by The Institute of Environmental Science and Research (ESR).

3.4.1. In 2019, 126.1 *Campylobacter* cases per 100,000 were reported in New Zealand. How many of these were estimated to be related to overseas travel?

Please express your answer as cases per 100,000.

3.4.2. Between 1998 and 2013 the annual number of yersiniosis notifications reported ranged between 383 and 546. Since 2015, the number of notifications for yersiniosis and the rate of yersiniosis notifications per 100,000 population has been increasing.

What was the yersiniosis notification rate per 100,000 reported in 2019?

Please express your answer as # of cases per 100,000.

Background: The two questions below are based on data presented by the Australian National Notifiable Disease Surveillance System (NNDSS) at the National Communicable Disease Surveillance Dashboard.

3.4.3. In the state of Queensland in Australia, 28 cases per 100,000 of cryptosporidiosis were reported in 2015. How many cases per 100,000 were reported in 2019?

Please express your answer as cases per 100,000.

3.4.4. In 2019, 657 cases of STEC were reported across all eight states in Australia. How many of these were reported in children aged 0-4 years?

Please express your answer as number of cases.

3.5 Questions based on disease surveillance data reported from Africa

3.5.1. Africa CDC publishes a "Weekly Event Based Surveillance Report". In the report from the last week of August 2023, it was stated that there had been 176,341 cases of cholera reported from 17 African Union (AU) member states since the beginning of 2023. Of these, 2,764 resulted in death, resulting in a case fatality rate (CFR) of 1.6%.

Thinking only of AU member states, what was the CFR of the country with highest rate?

Please express your answer as a percentage.

Use this equation:

- (cholera deaths / # cholera cases - in country with highest rate)*100

Background: The Institute of Health Metric and Evaluation (IHME) shares the data used for the Global Burden of Disease (GBD) study 2019. These data include estimates of diarrheal disease cases and deaths in the six WHO regions.

3.5.2. According to these data, mean death rates caused by rotavirus (12.4 deaths per 100,000 population) and *Shigella spp.* (9.3 deaths per 100,000 population) ranked highest among diarrheal disease causes in the WHO Africa region in 2019. What was the mean death rate for the diarrheal causative agent ranking 3rd?

Please express your answer as number of deaths per 100,000 population.

3.5.3. A systematic review and meta-analysis of the occurrence of intestinal parasites among food handlers of food service establishments in Ethiopia was recently published (in 2020). The meta-analysis included 20 studies published between 2001 and 2019 and the following parasites: Ascaris lumbricoides, Entamoeba histolytica/dispar, Taenia species, Hookworms, Giardia lamblia, Hymenlopsis nana, Strongloides stercolaris, Trichuris trichiura, Enterobius vermicularis, and Schistosoma mansoni.

What was the pooled prevalence estimate found among food handlers of food service establishments?

Please express your answer as a prevalence in percentage.

3.6 Questions based on disease surveillance data reported from the Middle East

Background: The Institute of Health Metric and Evaluation (IHME) shares the data used for the Global Burden of Disease (GBD) study 2019. These data include estimates of diarrheal disease cases and deaths in the six WHO regions.

3.6.1. According to these data, mean death rates caused by rotavirus (4.0 deaths per 100,000 population) ranked highest among diarrheal disease causes in the WHO Eastern Mediterranean region in 2019.

What was the mean death rate for *Campylobacter* in this region?

Please express your answer as a number of deaths per 100,000 population.

3.6.2. Across all countries in North Africa and the Middle East, the estimated mean death rate of cryptosporidiosis was 0.36 deaths per 100,000 population in 2019.

What was the death rate of cryptosporidiosis for the country with the highest rate?

Please express your answer as a number of deaths per 100,000 population.

4. Questions about access to improved water or improved sanitation

4.1. Unsafe water and sanitation and poor hygiene are, jointly, an important risk factor for mortality, particularly in Low- and Middle-Income Countries (LMIC). The World Bank publishes country-specific estimates on mortality rates attributed to unsafe water, unsafe sanitation and lack of hygiene (lack of WASH) (per 100,000 population).

Thinking of the countries in the region of South-east Asia (SEA), in 2016 what was the absolute difference between the countries with the lowest and the highest contributions of lack of WASH to their mortality rates? Express the difference in deaths per 100,000 population.

Express your answer in number per 100,000.

- Use the equation:
    - *Highest mortality due to lack of WASH in SEA 2016 (# per 100,000) − Lowest mortality due to lack of WASH SEA 2016 (# per 100,000)*

Background: The WHO/UNICEF Joint Monitoring Programme for Water Supply, Sanitation and Hygiene (JMP) has since 1990 reported estimates of progress on drinking water, sanitation and hygiene (WASH) at country, regional and global level. The following three question ask for estimates at the regional level using the eight Sustainable Development Goal (SDG) regions defined by the UN (Australia and New Zealand, Central and Southern Asia, Eastern and South-Eastern Asia, Europe and Northern America, Latin America and the Caribbean, Northern Africa and Western Asia, Oceania, Sub-Saharan Africa).

4.2 According to JMP, 54.7% of the rural population in Latin America and the Caribbean (LAC) had access to piped drinking water in 2000. For this region, what was the change in reported percentage from 2000 to 2019?

Please, express your answer in percentage change.

- Use this equation:
    - *((%rural piped drinking water 2019− %rural piped drinking water 2000) / %rural piped drinking water 2000)*100*

4.3 According to JMP, 12.43% of the population in the Oceanian region (i.e., excluding Australia and New Zealand) practiced open defecation in 2000. For this region, what was the reported change in the percentage of the population practicing open defecation from 2000 to 2019?

Please, express your answer in percentage change.

- Use this equation:
    - *((%population open defecation 2019 −%population open defecation 2000)/ %population open defecation 2000)*100*

4.4 Considering all the countries in the world in 2019, think of the country that the JPM reported as having the highest percentage of its population with no access to handwashing facilities at the household level. How much does this percentage differ between urban and rural areas in that country?

Please, express your answer in percentage difference.

- Use this equation:
    - *(% urban population with no handwashing – % rural population with no handwashing)*

5. Questions related to chemicals

5.1 According to the technical background report for the global mercury assessment 2018, global emissions of mercury (Hg) to air from different anthropogenic source sectors were estimated at 2200 tonnes in 2015.

What percentage was the result of emissions from stationary combustion of coal?

Please, express your answer as a percentage.

- Use this equation:
    - *(tonnes Hg from stationary combustion of coal / total tonnes Hg to air)*100*

5.2 Cadmium (Cd) is used as a core indicator to evaluate the status of the marine environment based on concentrations of Cd measured in seawater, biota and sediments. The Baltic Marine Environment Protection Commission, also known as the Helsinki Commission (HELCOM) report 2023 presents monitoring data collected in the Baltic Sea during the assessment period 2016 – 2021.

The evaluation of Cd concentrations in biota utilises fish muscle and liver (with no tissue conversion currently applied) of a range of fish species and soft tissues of mussels. According to the HELCOM 2023 report, the mean values of the Cd concentrations in biota ranged from 86 to 451 µg/kg across the 17 HELCOM sub-basins in the Baltic Sea.

What was the highest concentration found in biota in a single sample in any of these sub-basins?

Please, express your answer in µg/kg.

5.3 Polychlorinated biphenyls (PCBs) were produced in large quantities between the 1930s and 1980s and can still be found everywhere. The Stockholm Convention obliges countries to eliminate the use of PCBs in equipment by 2025 and make determined efforts to employ the environmentally sound management of waste liquids and equipment contaminated with PCBs by 2028. By August 2023, 165 countries had implemented National Implementations Plans to reach the 2028 global target of the elimination of 14,027 tons PCBs.

What is the target for the African region expressed in percentage of the global target?

Please, express your answer in percentage point.

5.4 The production of chlorofluorocarbons (CFCs) that would ultimately be released to the atmosphere was banned globally in 2010 under the Montreal Protocol. Measurements and modelling have been used to show how atmospheric emissions of five specific CFC types (xCFC-113,CFC-113a,CFC-114a,CFC-115, CFC-115a) changed between 2010 and 2020.

What was the 2020 total annual emission rate (i.e., weight/year) of these five CFCs, expressed in percentage terms relative to 2010 emission rate? Technically, CFC weight is expressed as ODP-Gg yr-1, where ODP = CFC-11-equivalent ozone-depleting potential.

Please, express your answer as a percentage of 2010 emission.

- Use this equation:
    - *(CFCs emission 2020/CFCs emission 2010) *100*

5.5 The WHO recommended permissible limit for Pb in the consumption of fish to be 0.2 mg/kg. In a study of heavy metals in sediments and in Common Carp from a reservoir in Lesotho, the mean concentration level of Pb in sediments at one sampling site (of three) was 0.25 mg/kg.

What was the mean concentration of Pb in gills of Common carp from the same sampling site, in mg/kg?

Please, express your answer in mg/kg.

5.6 The AflaCohort Birth Cohort Study (2015–2019) was conducted in Banke, a tropical district in the southern plains of Nepal. A rolling recruitment strategy enrolled 1,675 healthy pregnant women and a total of 1,650 gestational serum samples were analysed for AFB1-lys adducts, an established biomarker of dietary aflatoxin exposure over the previous 2–3 months. Weekly consumption frequencies included 31% of the cohort reporting consumption of groundnuts, 3% reporting maize, and 2% reporting both groundnuts and maize. The mean AFB1-lys adduct level in those who did not eat maize or groundnuts in the past week was 2.4 pg/mg albumin adducts.

What was the mean level in pregnant women who consumed both maize and groundnut in the previous week?

Please, express your answer in pg/mg.

5.7 Aflatoxin M1 (AF M1) is a hydroxylated metabolite of aflatoxin B1 which may be present in human milk from exposed mothers and could pose a risk to neonates. In a study in Turkey, published in 2010, 75 breast milk samples tested positive for aflatoxin M1 by high performance liquid chromatography HPLC.

What percentage of these samples had AF M1 concentrations exceeding 100 ng/l?

Please, express your answer as a percentage.

5.8 What was the annual number of deaths from mesothelioma caused by asbestos exposure in the UK in 2021 as reported in the Asbestos Health Statistics by the UK Asbestos Training Association (UKATA)?

Please, express your answer as number of deaths.

5.9 From a major study on the effect of urban air pollution on lung cancer incidence in Sweden, published in 1994, what was the ratio of the estimated number of lung cancers per year caused by inhaled arsenic to the number caused by nickel?

Please, express your answer as a rate.

- Use this equation:
    - *# cancer cases per 1,000,000 per year from arsenic /# cancer cases per 1,000,000 per year from nickel*

5.10 A study (published in 1998) conducted in a region in Chile with of around 440,000 inhabitants from 1989-1993, investigated the mortality after exposure to arsenic-contaminated drinking water with an average 5-year exposure ranging from 0.043 (in 1990-94) to 0.569 (in 1955-59) mg/l over a period of 45 years (1950-94). Standardized mortality ratios were estimated by dividing observed deaths by expected deaths. The standardized mortality ratio (SMR) in men for bladder cancer was 6, kidney cancer was 1.6, and skin cancer was 7.7.

What was the SMR for lung cancer in men?

Please, express your answer as a standardized mortality rate (SMR).

5.11 In a study, published in 1997, of cerebral infarction in those exposed to arsenic by drinking well water, the multivariate-adjusted odds ratio was 6.9 for those that drank water with an arsenic concentration greater than 0.3 mg/l and 4.5 for those who consumed well water with an arsenic content of 0.0501 - 0.2999 mg/l.

What was the odds ratio for those in the low exposure group (0.0001 - 0.05 mg/l)?

Please, express your answer as an odds ratio (OR).

5.12 An important food source of human dioxins exposure are pork products. In the Netherlands, pork products are frequently tested by the government and food business operators for the presence of dioxins; samples are first analysed using a CALUX (Chemical Activated Luciferase gene eXpression) test, which typically cost about €280 per test (in 2022).

If the CALUX test was classified “suspect”, what was the cost of a follow-up gas chromatography–mass spectrometry confirmatory test, in € (euros)?

5.13 In a case-control study, whose results were published in 2019, concentrations of 49 persistent organic compounds (POPs) were measured in both adipose tissue and serum samples from breast cancer patients who underwent partial or total mastectomies, lymph node biopsies and sampling of the adipocytic tumor microenvironment. Adjusted, unconditional logistic models were used to study associations between POP concentrations and risk of metastasis and other hallmarks of cancer aggressiveness, assessed for two sub-population groups with BMI ≥25 kg/m^2 (n = 44) and BMI <25 kg/m^2 (n = 47). The models were adjusted for age, smoking, body mass index, menopause and familial history of breast cancer. Findings were expressed in terms of odds ratios (OR).

Tetrachlorodibenzo-p-dioxin (2.3.7.8-TCDD) concentrations in adipose tissue are positively associated with the risk of metastasis in patients with BMIs ≥25 kg/m^2.

What was the corresponding mean OR value? Please, express your answer as an OR.

5.14 The Toxics Release Inventory (TRI) under the US Environmental Protection Agency is a resource for learning about toxic chemical releases and pollution prevention activities reported by industrial and federal facilities. In 2021, 799 facilities submitted a TRI form reporting on the facility's dioxin releases.

According to these reports, how large a percentage of dioxin releases were disposed off site, primarily in landfills?

Please, express your answer as a percentage.

5.15 In a review study from Germany published in 2018 sources for environmental contamination of PCDD/Fs and PCBs relevant to food safety were investigated. The study concluded that a daily dioxi-like(dl)-PCB intake for suckler cow herds must in average be less than 2 ng PCB-TEQ/day for the meat not to exceed the EU regulatory limit.

To what maximum concentration in grass expressed as ng PCB-TEQ/kg dm (dry matter) does this value translate?

Please, express your answer in ng PCB-TEQ(toxic equivalents)/kg dm (dry matter).

5.16 Over the past few decades, the Japanese Ministry of the Environment has been biomonitoring dioxins in the general Japanese population and has taken measures to reduce dioxin exposure. A study published in 2019, compared blood dioxin levels collected in two surveys conducted in the Japanese population in 2002–2010 and 2011-2016, respectively.

What was the percentage change in the median blood dioxin level from the first to the second survey?

Please express your answer as a percentage.

- Use this equation
- *((median blood dioxin level 2011-2016 – median blood dioxin level 2002-2010)/ median blood dioxin level 2002-2010)*100*

5.17 A global systematic review published in 2022 extracted data on the concentration of aflatoxins in different types of nuts. Scientific databases were searched systematically from 2000 to 2020 and ended up including 73 studies. Based on the results, the concentration of aflatoxin B1 (AFB1) was found to be highest in peanuts with a mean concentration across all studies on 32.82 µg/kg.

What was the mean concentration of AFB1 found in peanuts in the country with the highest mean concentration across all studies?

Please express your answer in µg/kg.

5.18 Liver injury and hepatocellular carcinoma (HCC), one of the major types of liver cancer, are considered the main toxic impact of Aflatoxin B1 (AFB1). Worldwide, approximately 5–28% of HCC occurrences are attributed to Aflatoxin exposure.

In 2020, what was the global estimated crude incidence rate (i.e., the incidence rate calculated without considering possible confounding factors) of liver cancers across both sexes and all age groups?

Please express your answer as number of cases per 100,000 population.

5.19 In a study published in 2019, aflatoxin M1 (AFM1) contamination levels in raw milk samples collected from Punjab, Pakistan, were investigated. A total of 960 milk samples from five different regions were collected every month in 2015. The AFM1 level in raw milk was analysed by the ELISA technique.

How large a percentage of the samples exceeded the United States permissible maximum residue limits (MRL) of 0.50 µg/L?

Please express your answer as a percentage.

Table B3. Elicitation videos (with links).

Title	link
Intro to the SEJ study	https://youtu.be/IdY2SXMrnHU
Intro to the Classical Model for Structured Expert Judgment	https://youtu.be/b8uMlhGeq4w
Uncertainty quantification	https://player.hihaho.com/c3cd2aa6-0085-4a59-9172-e9e9cc35b0b4
Interactive Q&A session	https://youtu.be/PJs43dpsp1Y?si=YkmvllnkiX3pUVzM

Table B4. List of training questions

Appendix C Elicitors and Experts

Table C1. Criteria for selection of the experts.

Criteria	**Academic profile**	**Non-academic (governmental) profile**
Relevance of university degree attained*	None (2 points); BSc or equivalent (4 points); MSc or equivalent (6 points); PhD/doctoral or equivalent (8 points); postdoctoral or equivalent (10 points)	None (2 points); BSc or equivalent (4 points); MSc or equivalent (6 points); PhD/doctoral or equivalent (8 points); postdoctoral or equivalent (10 points)
Years of experience on call topic	1-3 (2 points); 3-6 (4 points); 6-10 (6 points); 10-15 (8 points); >15 (10 points)	1-3 (2 points); 3-6 (4 points); 6-10 (6 points); 10-15 (8 points); >15 (10 points)
Relevant publications or reports on corresponding hazard(s) and region(s)	<10 (2 points); 11-20 (4 points); 21-30 (6 points); 31-40 (8 points); >40 (10 points)	Not applicable
Level of seniority (position held)	Not applicable	Trainee/intern (2 points); junior/assistant (4 points); officer/investigator (6 points); senior officer/PI (8 points); director/head of lab/service (10 points)

*Experts holding a university degree below the BSc level or its equivalent (e.g., undergraduate students) were excluded from consideration.

Table C2. List of active elicitors

Elicitor
Eduard Grau-Noguer
Zoe Baldwin
Emi Grace Mary Gowshika
Uswatun Hasanah
Stanley Chen
Emrecan Özeler
Muhammad Tanveer Munir
Abiodun Folake Abiola Omogoye
Reha Onur Azuziglu
Stephanie Poling
Miranda Nonikashvili
Eiki Yamasaki
Janet Rymound
Lisa O'Connor
Pankaj Dhaka
Ankur Aggarwal
Sara Faife
Maria Olorunsola
Sarah Hagan
Devin LaPolt

Jamila Seaton
Dhanalakshmi Marimuthu
Dikshit Poudel
Iqra Zaheer
Selam Alemu
Kossi Brice Boris
Nada Alasiri
Ana Margarida Pignateli Vasconcelos de Alho
Alessandra Primavera
Lapo Mughini-Gras
Maria Francesca Iulietto
Malak Elbassuny
Tina Nane

Table C3. List of experts whose assessments were used in the analysis.

Musa Imam	Abubakar
Dayo	Adeyemo
Hanaa	Al Enizi
Fadi	Al natour
Nada	Alasiri
Ana Margarida	Alho
Kebede	Amenu
Mohammed Badrul	Amin

Ljupcho	Angelovski
Gabriela	Arrifano
William	Arthur Petri
HUIHUI	BAO
Alessandra	Barlaam
Norman	Beatty
Dawn	Blackburn
Beau	Bruce
Daudet	Byakya Kikukama
Ana Karina	Carrascal Camacho
Adriano	Casulli
Laurie	Chan
Stanley	Chen
José	Chen-Xu
Roger	Cooke
Mariana Oliviera	Corda
Maria Elena	Crespo López
Premanshu	Dandapat
Marthe	De Boevre
Mateus	de Souza Ribeiro Mioni
Ken	Diplock
Polikseni	Drazho
Ayebare	Dreck
Dwiyitno	Dwiyitno
Mariam	Elkhayat
Mariem	Ellouze
Ayman	El-Shibiny
Olanrewaju	Fayemi
Luria	Founou
Eelco	Franz
Paula	Fujiwara
Abhishek	Gautam

Herman	Gibb
David	Goldman
Jorge	Gomez-Marin
Wiem	Guissouma
Pradip	Gyawali
Jisun	Haan
Theoneste	Hagenimana
Ekhlas	Hailat
Abdulsamie	Hanano
Abul	Hashem
Jean Paul	Hategekimana
Nicola	Holden
Darren	Holland
Kristy	Hope
Paul	Hunter
Christelle	Iskandar
Sanu	Jacob
Yasmin	Jahan
Ole	Jakob
Elizabeth Alejandra	Jara Torres
Abdurrahman	Jibril
Anita	Kambhampati
Layal	Karam
Adrew	Karasick
Jongsoo	Kim
Erica	Kintz
Ivana	Klun
Pauline	Kooh
Chihaya	Koriyama
Brian	Lassen
Francoise, Soizick	Le Guyader
Salvador,	Liliana
Peter	Lindberg Nejsum

Naeemah	Logan
HAMDI	Lotfi
Jingrang	Lu
Daniel	Lucas
Sara	Lupton
Shannon	Majowicz
Jenny	Maloney
Carla	Martins
Alison	Mather
Zaffar	Mehmood
Augustin Octavian	Mihalache
Naim	Montazeri
Peter	Moono
Patricia Carolina	Moyano
Lapo	Mughini-Gras
Adrian	Muwonge
Phoebe	Nabunya
Nachimata	Nambela
Neda	Nasheri
Luis Augusto	Nero
Subhaprada	Nishtala
Patrick	Njage
Francisco	Olea Popelka
Bukola	Onarinde
Helen	Onyeaka
Ynes	Ortega
Himadri	Pal
Vijay	Pal
Katherine	Paphitis
Julio	Parra-Flores
Arti	Pillay
Marina	Pinheiro
Uelinton	Pinto

Thelma Veronica	Poggio
Novalia	Rachmawati
Voniarisoa Razafindramary	Rahanjavelo
Tulsi	Ram Gompo
Deepak	Rawool
Lucy	Robertson
Deyci	Rodriguez
Fernando	Rosado Spilki
Mirosław	Różycki
Karina	Saadi Siú
Maria	Saldias Molina
Adil	Salman
Fernando	Sampedro
Craig	Shadbolt
Shamsi	Shokoofeh
Maansi	Shukla
Karen	Signori Pereira
Vijay Pal	Singh
Tuti	Siregar
Caroline	Smith DeWaal
Lama	Soubra
Judi	Spungen
Jaya	Sundaram
Nimisha	Suraj
Pondpan	Suwathada
Noriyuki	Suzuki
Nao	Takeuchi-Storm
Sayed	Tariq Pachakhan
Nozomi	Tatsuta
Kate	Thomas
Michael	Tomori
Eduardo Cesar	Tondo

Radestya	Triwibowo
Tomoaki	Tsutsumi
Helena	Ullyartha Pangaribuan
Aleksandra	Uzelac
Abimbola	Uzomah
Ivar	Vågsholm
Nicolas	Valdivieso Cariola
Helga	Waap
Timothy	Wade
Yibaina	Wang
Daniel	Weller
Anthony	Wilson
Felicia	Wu
Ricardo	Yamatogi
Ian	Young
Amona	Yousif Hamed
Hamdouni	Youssef
Lei	Zhang
Shugufta	Zubair
Paulina	Zurita Urrea

Appendix D Mathematical validation and aggregation

The Classical Model for Structured Expert Judgement (SEJ) proposes a validation framework to aggregate experts' assessments. The framework relies on the set of calibration questions and two objective measures of performance to derive aggregation weighting schemes. There are two generic, quantitative measures of expert performance, referred to as *statistical accuracy* (also referred to as calibration score) and *information*. Statistical accuracy is also termed "calibration" in older literature. Loosely, statistical accuracy measures the statistical likelihood that a set of experimental results correspond, in a statistical sense, with an expert's assessments. More precisely, under the Classical Model, statistical accuracy is scored as the p-value at which we would falsely reject the hypothesis that expert's probability statements were statistically accurate. In this study the 5^{th}, 50^{th} and 95^{th} percentiles, or quantiles, were elicited from each expert for each of the continuous variables. Hence, the range of possible outcomes of each variable can be divided into 4 intervals: less than or equal to the 5^{th} percentile, greater than the 5^{th} percentile value and less than or equal to the 50^{th} percentile value, etc. The probabilities for these intervals are expressed as a vector

$p = (p_1, p_2, p_3, p_4) = (0.05, 0.45, 0.45, 0.05)$.

Statistical Accuracy

If N variables are assessed, each expert may be regarded as a statistical hypothesis, namely that each realization falls in one of the four inter-quantile intervals with probability vector p. Suppose we have realizations $x_1,\ldots,x_N$ of these quantities. We may then form the sample distribution of the expert's inter quantile intervals as:

$$s_1(e) = \#\{\, i \mid x_i \le 5\text{th percentile}\}/N$$
$$s_2(e) = \#\{\, i \mid 5\text{th percentile} < x_i \le 50\text{th percentile}\}/N$$
$$s_3(e) = \#\{\, i \mid 50\text{th percentile} < x_i \le 95\text{th percentile}\}/N$$
$$s_4(e) = \#\{\, i \mid 95\text{th percentile} < x_i \}/N$$
$$s(e) = (s_1,..,s_4)$$

Note that the sample distribution depends on the expert e. If the realizations are indeed drawn independently from a distribution with quantiles as stated by the expert, then the quantity

$$2NI(s(e) \mid p) = 2N \Sigma_{i=1..4}\, s_i ln(s_i / p_i) \qquad (1)$$

is asymptotically distributed as a Chi-square variable with 3 degrees of freedom. This is the likelihood ratio statistic, and I(s | p) is the relative information of distribution s with respect to p. Extracting the leading term of the logarithm yields the familiar chi-square test statistic for goodness of fit.

If after a few realizations the expert were to see that all realization fell outside his 90% central confidence intervals, (s)he might conclude that these intervals were too narrow and might broaden them on subsequent assessments. This means that for this expert the uncertainty distributions are *not* independent, and (s)he learns from the realizations. Expert learning is not a goal of an expert judgment study. Rather, the problem owner wants experts who do not need to learn from the elicitation. Independence is not an assumption about the expert's distribution but a desideratum of the problem owner. Hence the decision maker (see below) scores expert e as the statistical likelihood of the hypothesis

H_e: "the inter quantile interval containing the true value for each variable is drawn independently from probability vector p."

A simple test for this hypothesis uses the test statistic (1), and the likelihood, or p-value, for testing the hypothesis that the expert is *statistically accurate:*

$$SA(e) = \text{p-value}(e) = \text{Prob}\{2NI(s(e)\ p) \ge r \mid H_e\}$$

where r is the value of (1) based on the observed values $x_1,\ldots x_N$. It is the probability under hypothesis H_e that a deviation at least as great as r should be observed on N realizations if H_e were true. *SA* scores are dimensionless and can be compared across studies. However it is appropriate to compare different hypothesis tests (across different studies, for example) by equalizing the effective number of realizations. To compare scores on two data sets with N and N' *realizations*, we simply use the minimum of N and N' in (1), without changing the sample distribution s.

The *Statistical Accuracy* score is sometimes termed the *Calibration Score.* Although this score uses the language of simple hypothesis testing, it must be emphasized that we are not rejecting expert-hypotheses; rather we are using this language to measure the degree to which the data supports the hypothesis that the expert's probabilities are accurate. Low scores, near zero, mean that it is unlikely that the expert's probabilities are correct. High scores, near 1, indicate good support.

Information

The second scoring variable is information. Loosely, the information in a distribution is the degree to which the distribution is concentrated. Information cannot be measured absolutely, but only with respect to a background measure. Being concentrated or "spread out" is measured relative to some other distribution. Commonly, the uniform and log-uniform background measures are used. Measures which are not relative in this sense, such as the standard deviation or probability interval, inherit the physical dimension of the underlying variable (meters, micro grams per cubic meter, etc) and cannot be compared across variables with physical different dimensions.

Measuring information requires associating a density with each assessment of each expert. To do this, we use the unique density that complies with the experts' quantiles and is minimally informative with respect to the background measure. This density can easily be found with the method of Lagrange multipliers. For a uniform background measure, the density is constant between the assessed quantiles. The background measure is not elicited from experts as indeed it must be the same for all experts; instead, it is chosen by the analyst based on all expert assessments.

The uniform and log-uniform background measures require an *intrinsic range* on which these measures are concentrated. The classical model implements the so-called *k*% overshoot rule: for each item we consider the smallest interval *I = [L, U]* containing all the assessed quantiles of all experts and the realization, if known. This interval is extended to

$$I^* = [L^*, U^*]; \text{ where } L^* = L - k(U-L)/100;\ U^* = U + k(U-L)/100.$$

The value of k is chosen by the analyst. A large value of *k* tends to make all experts look quite informative and tends to suppress the relative differences in information scores. *The* **information score** of expert *e* on assessments for uncertain quantities 1…N is

$$\text{Inf (e) =Average Relative information w.r.t. Background} = (1/N) \sum_{i = 1..N} I(f_{e,i} \mid g_i)$$

where g_i is the background density for variable *i* and $f_{e,i}$ is expert *e*'s density for item i. This is proportional to the relative information of the expert's joint distribution given the background, under the assumption that the variables are independent. As with statistical accuracy, the assumption of independence here reflects a desideratum of the decision maker and not an elicited feature of the expert's joint distribution. The information score does not depend on the realizations. An expert can give her/himself a high information score by choosing his quantiles very close together. The information score of *e* depends on the intrinsic range and on the assessments of the other experts. Hence, information scores cannot be compared across studies.

The above information score is chosen because it is familiar, tail insensitive, scale invariant and "slow". The latter property means that relative information is a slow function; large changes in the expert assessments produce only modest changes in the information score. This contrasts with the likelihood function in the statistical accuracy score, which is a very "fast" function. This causes the product of statistical accuracy and information to be driven by the statistical accuracy score. These two performance measures are negatively correlated in many expert data sets, that is, experts with high information tend to have lower statistical accuracy (see section ****).

Combination: Decision Maker

Combining experts should *always* be applied to experts' densities as described above and *NOT* to the experts' quantiles themselves. Simply averaging quantiles is known to produce overconfident results, see link (5) above.

The simplest combination scheme simply averages the experts' densities with equal weighting (EWDM). Combinations based on expert performance use the dimensionless combined *score* of expert *e* which serves as an (unnormalized) weight for *e*:

$$w_\alpha(e) = SA\ (e) \times Inf\ (e) \times \mathbb{1}_\alpha(SA(e) \geq \alpha), \qquad (2)$$

where $\mathbb{1}_\alpha(SA(e)\alpha) = 1$ if $SA(e) \geq \alpha$, and is zero otherwise. The combined score thus depends on α; if $SA(e)$ falls below cut-off level α, expert *e* is unweighted. The presence of a cut-off level is imposed by the requirement that the combined score be an asymptotically strictly proper scoring rule. That is, an expert maximizes his/her long run expected score by and only by ensuring that his probabilities $p = (0.05, 0.45, 0.45, 0.05)$ correspond to his true beliefs. α is similar to a significance level in simple hypothesis testing, but its origin is to measure 'goodness' and not to reject hypotheses. When presenting data of combined scores, one typically sets $\alpha = 0$, so that the combined score is simply $SA(e) \times Inf(e)$.

A combination of expert assessments is called a "decision maker" (DM). All decision makers discussed here are examples of linear pooling; the classical model is essentially a method for deriving weights in a linear pool. "Good expertise" corresponds to good statistical accuracy (high statistical likelihood, high p-value) and high information. Weights that reward good expertise and pass these virtues on to the decision maker are desired.

The reward aspect of weights is very important. We could simply solve the following optimization problem: find a set of weights such that the linear pool under these weights maximizes the product of statistical accuracy and information. Solving this problem on real data, one finds that the weights do not generally reflect the performance of the individual experts. As an expert's influence on the decision maker should not appear haphazard, and "gaming" the system with assessments tilted to achieve a desired outcome should be discouraged, we must impose a strictly scoring rule constraint on the weighting scheme.

The scoring rule constraint requires the term $(SA(e) \geq \alpha)$ in eq (2), but does not indicate what value of α we should choose. Therefore, we choose α to maximize the combined score of the resulting decision maker. Let $DM_\alpha(i)$ be the result of linear pooling for any item *i* with weights proportional to (2):

$$DM_\alpha(i) = \sum_{e=1,..E} w_\alpha(e)\, f_{e,i} \ / \sum_{e=1,..E} w_\alpha(e) \qquad (3)$$

Model evaluation

The optimized global Performance Weighted DM (PWDM_Opt) is DM_{α^*} where α* maximizes

$$SA\ score(DM_{\alpha^*}) \times information\ score(DM_{\alpha^*}). \qquad (4)$$

The *Performance Weighted global DM non-Optimized (*PWDM*)* is is given by (4) with $\alpha = 0$.

These DMs are termed "global" since the information score is based on all the assessed calibration variables.

A variation on this scheme allows a different set of weights to be used for each item. This is accomplished by using information scores for each item rather than the average information score:

$$w_\alpha\,(e,i) = 1_\alpha(SA\ score) \times SA\ score(e) \times I(f_{e,i} \mid g_i) \qquad (5)$$

For each α we define the *Item Weight* DM_α for item *i* as

$$IDM_\alpha(i) = \sum_{e=1,..E} w_\alpha(e,i)\, f_{e,i} \ / \sum_{e=1,..E} w_\alpha(e,i) \qquad (6)$$

The optimized Item Weight DM (IWDM_Opt) is IDM_{α^*} where α^* maximizes

$$SA\ score(IDM_{\alpha^*}) \times information\ score(IDM_{\alpha^*}). \quad (7)$$

The Item weight DM non-Optimized (IWDM) I is given by (7) with $\alpha = 0$.

Item weights are potentially more attractive as they allow an expert to up- or down- weight her/himself for individual items according to how much (s)he feels (s)he knows about that item. "Knowing less" means choosing quantiles farther apart and lowering the information score for that item. Of course, good performance of item weights requires that experts successfully perform this up-down weighting. Anecdotal evidence suggests that item weights improve over global weights as the experts receive more training in probabilistic assessment.

Both item and global weights can be described as optimal weights under a strictly proper scoring rule constraint. With both global and item weights, statistical accuracy strongly dominates over information, and information serves to modulate between more or less equally well calibrated experts.

Since any combination of expert distributions yields assessments for the calibration variables, any combination can be evaluated on the calibration variables. In particular, we can compute the statistical accuracy and the information of any proposed decision maker. We should hope that the decision maker would perform better than the result of simple averaging (*EWDM*). The global and item weight *DM*'s discussed above (optimized or not) are *Performance based DM's. DM*'s are evaluated in-sample by treating the *DM* as a new expert and scoring its performance on the calibration variables. Out-of-sample arises when the *DM* is initialized on a subset of calibration variables and scored on the complementary subset.

An evaluation of expert performance in the previous WHO study is found in Aspinall et al. (2016). The performance based DM's have been shown to be superior to equal weighting both in-sample (link 3) and out-of-sample (link 4) and the expert performance has been shown to be a persistent property of experts (link 2).

Appendix E

- Regional/economical/global proxies