











## ORIGINAL ARTICLE OPEN ACCESS

# Evaluating the Performance of Agreement Metrics in a Delphi Study on Chemical, Biological, Radiological and Nuclear Major Incidents Preparedness Using Classical and Machine Learning Approaches

Hassan Farhat<sup>1,2</sup>  | Alan M. Batt<sup>3,4</sup>  | Mariana Helou<sup>5,6</sup>  | Heejun Shin<sup>7,8,9,10</sup> | James Laughton<sup>2</sup>  | Carolyn Dumbleck<sup>11</sup> | Arezoo Dehghani<sup>12,13</sup> | Fatemeh Rezaei<sup>13</sup>  | Nidaa Bajow<sup>14</sup>  | Luc Mortelmans<sup>15,16,17</sup>  | Walid Abougalala<sup>18</sup> | Roberto Mugavero<sup>19,20,21</sup> | Gregory Ciottoni<sup>10,22</sup>  | Guillaume Alinier<sup>2,23,24,25</sup>  | Mohamed Ben Dhiab<sup>1</sup> 

<sup>1</sup>Faculty of Medicine “Ibn El Jazzar,” University of Sousse, Sousse, Tunisia | <sup>2</sup>Ambulance Service, Hamad Medical Corporation, Doha, Qatar | <sup>3</sup>Queen's University, Kingston, Ontario, Canada | <sup>4</sup>Monash University, Melbourne, Victoria, Australia | <sup>5</sup>School of Medicine, Lebanese American University, Beirut, Lebanon | <sup>6</sup>Lebanese American University-Rizk Hospital, Beirut, Lebanon | <sup>7</sup>Soonchunhyang Disaster Medicine Center, Bucheon, South Korea | <sup>8</sup>Soonchunhyang University Bucheon Hospital, Bucheon, South Korea | <sup>9</sup>Shin's Disaster Medicine Academy, Seoul, South Korea | <sup>10</sup>Harvard Medical School, Harvard University, Cambridge, Massachusetts, USA | <sup>11</sup>Department of Disaster Management, Alberta Health Services | <sup>12</sup>Safety Promotion and Injury Prevention Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran | <sup>13</sup>Department of Health in Emergencies and Disasters, Health Management and Economic Research Center, Isfahan University of Medical Sciences, Isfahan, Iran | <sup>14</sup>Security Forces Hospital, Riyadh, Saudi Arabia | <sup>15</sup>European Society for Emergency Medicine, Belgium | <sup>16</sup>Catholic University of Leuven, Belgium | <sup>17</sup>Free University Brussels, Belgium | <sup>18</sup>Hamad Medical Corporation, Doha, Qatar | <sup>19</sup>Department of Electronic Engineering – DIE, University of Rome “Tor Vergata, Rome, Italy | <sup>20</sup>Centre for Security Studies – CUFSS, University of the Republic of San Marino, San Marino | <sup>21</sup>Observatory on Security and CBRNe Defense – OSDIFE | <sup>22</sup>Harvard T.H. Chan School of Public Health, USA | <sup>23</sup>School of Health and Social Work, University of Hertfordshire, Hatfield, UK | <sup>24</sup>Weill Cornell Medicine-Qatar, Doha, Qatar | <sup>25</sup>Faculty of Health and Life Sciences, Northumbria University, Newcastle upon Tyne, UK

**Correspondence:** Hassan Farhat ([hassen.farhat@gmail.com](mailto:hassen.farhat@gmail.com))

**Received:** 9 December 2024 | **Revised:** 24 March 2025 | **Accepted:** 24 March 2025

**Funding:** The authors received no specific funding for this study.

**Keywords:** agreement analysis | Delphi study | disaster medicine | expert's opinion | MENA

## ABSTRACT

Delphi studies in disaster medicine lack consensus on expert agreement metrics. This study examined various metrics using a Delphi study on chemical, biological, radiological, and nuclear (CBRN) preparedness in the Middle East and North Africa region. Forty international disaster medicine experts evaluated 133 items across ten CBRN Preparedness Assessment Tool themes using a 5-point Likert scale. Agreement was measured using Kendall's W, Intraclass Correlation Coefficient, and Cohen's Kappa. Statistical and machine learning techniques compared metric performance. The overall agreement mean score was  $4.91 \pm 0.71$ , with 89.21% average agreement. Kappa emerged as the most sensitive metric in statistical and machine learning analyses, with a feature importance score of 168.32. The Kappa coefficient showed variations across CBRN PAT themes, including medical protocols, logistics, and infrastructure. The integrated statistical and machine learning approach provides a promising method for understanding expert consensus in disaster preparedness, with potential for future refinement by incorporating additional contextual factors.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2025 The Author(s). *Journal of Contingencies and Crisis Management* published by John Wiley & Sons Ltd.

## 1 | Introduction

The use of Delphi studies in disaster medicine research is increasing (Gray et al. 2023; Hill et al. 2022; Niederberger and Spranger 2020). These studies offer a structured approach to exploring expert knowledge and achieving consensus on critical disaster preparedness and response. They achieve this by collecting and aggregating informed judgments from a panel of experts on specific topics. These studies are advantageous for exploring problems where experimental evidence is often limited and expert opinion plays a crucial role in decision-making and policy creation (Keating and Hanger-Kopp 2020).

The benefits of Delphi studies in disaster medicine include gathering geographically dispersed experts, maintaining anonymity to reduce bias and facilitating iterative feedback to refine opinions (Alammary 2022; Munblit et al. 2022). Delphi studies can also help identify priorities, develop guidelines, and forecast future trends in disaster management. However, criticisms of the method have also been noted, such as researcher bias in panel selection may skew expert opinions toward certain institutional priorities, while participant fatigue can result in premature or superficial consensus on critical preparedness issues (Banno et al. 2020; Spranger et al. 2022). Previous studies have emphasised the importance of expert consensus in shaping public health responses to CBRN terrorism in the MENA region (Mani et al. 2024). Their findings highlight how regional disparities in healthcare infrastructure and preparedness capabilities can influence expert agreement, highlighting the need for metrics like Kappa to account for contextual variations. Furthermore, these studies suggest that integrating standardised metrics with localised assessments could enhance the development of actionable, region-specific protocols for CBRN preparedness (Uuk et al. 2024).

Yet, expert consensus in chemical, biological, radiological, and nuclear (CBRN) preparedness elements is crucial to enhancing readiness and response capabilities, especially in regions with diverse healthcare systems and limited empirical data, such as the Middle East and North Africa (MENA) region. Such consensus could facilitate the development of standardised protocols and training programs to fulfil regional needs. Aggregating expert knowledge helps identify critical gaps in preparedness, prioritise interventions, and establish best practices that can be adapted across different contexts (Ranse et al. 2024). Experts' agreement also facilitates policy-making and guides the building of impactful preparedness measures. In the MENA region, where CBRN threats may vary significantly between countries, expert consensus helps create a unified approach while accounting for local variations in terms of potential risks due to industrial sectors, conflicts, or terrorist threats. The Delphi collaborative approach improves the overall quality of preparedness plans and facilitates international cooperation and knowledge sharing.

One critical aspect of Delphi studies (among several, such as the definition of expert, timeframes, and planning of logistics) is the selection of appropriate metrics to measure agreement among experts effectively. The choice of agreement measurement metrics can significantly impact the interpretation of results and conclusions. Common metrics used in Delphi studies include

central tendency and dispersion measures, such as mean, median, standard deviation, and interquartile ranges (IQR) (Franc et al. 2023). Many studies have also employed more advanced statistical agreement measures, such as Kendall's W coefficient of concordance, intraclass correlation coefficient (ICC), and Cohen's of Kappa, among other metrics (Hoenig et al. 2024; Peng et al. 2024).

Despite the widespread use of Delphi studies in disaster medicine, there remains a lack of consensus on the most appropriate metrics for measuring expert agreement. This study aims to explore various agreement metrics and determine the most effective consensus measurement metric using the results of a Delphi study that assessed experts' opinions on preparedness for CBRN emergencies in the MENA region.

## 2 | Method

### 2.1 | Study Design and Setting

This quantitative analysis used the data set of the outcome of a previously published cross-sectional study. In this study, international experts in disaster medicine were invited to participate in an online Delphi panel using the Phonic® web application to seek consensus on a health Preparedness Assessment Tool (PAT) for CBRN incidents. This cross-sectional study used validated 5-point Likert scale items where the detailed outcomes were explained in a previously published study (Farhat et al. 2024). It showed high agreement among panellists about the proposed operational flowcharts, assessment tools, and training scenarios for CBRN incidents. It also revealed four distinct clusters among the experts' consensus data, emphasising European experts' response engagement (Farhat et al. 2024).

This study was approved by the Ethical Committees of the Faculty of Medicine "Ibn Eljazzar" of Sousse in Tunisia and Hamad Medical Corporation's Medical Research Center in Qatar (references CEFMS 110/2022 and MRC-01-22-258, respectively).

### 2.2 | Participants and Sampling

Forty multidisciplinary, international specialised experts in disaster medicine, with particular expertise in CBRN, participated in the online Delphi study between 1 November 2023 and 30 January 2024. The study targeted experts in disaster medicine with academic qualifications, publications, and interests in disaster medicine research in the MENA region. Purposive sampling, enhanced by snowball sampling, was utilised in this study.

### 2.3 | Variables

Various metrics were deployed to assess expert agreement, each rated on a 5-point Likert scale. Primary variables included the raw Likert responses (experts' responses) and four agreement measures: the percentage of agreement (PA), Kendall's W

Coefficient of concordance, Intraclass Correlation Coefficient (ICC) and Cohen's Kappa. The PA represents the proportion of responses for each of the scales (poor, fair, moderate, good and excellent). The PA ranges from zero to 100% (Drumm et al. 2022). Kendall's W assessed overall agreement among raters, ranging from 0 (no agreement) to 1 (complete agreement) (Denham 2016). The ICC measured rating reliability, ranging from 0 to 1 (Denham 2016). The ICC accounts for correlation and agreement between measurements. Cohen's Kappa coefficient evaluates inter-rater agreement while adjusting for chance (Vergni et al. 2021). It ranges from  $-1$  to  $1$ , where  $1$  indicates perfect agreement. Additionally, a categorical indicator of the agreement was created for the PA (0–20%: “Poor”, 21–40%: “Fair”, 41–60%: “Moderate”, 61–80%: “Good” and 81–100%: “Excellent”) to facilitate the machine learning analysis.

The PAT comprised ten general metrics or themes, including 48 specific metrics, further detailed into 133 individual metrics or items (Supporting Information: Appendix 1). The ten general metrics are: 1) Medical Protocols and Logistics, 2) Infrastructure readiness for CBRN Incidents in the MENA region, 3) Decontamination capabilities, 4) Specialised human resources capabilities, 5) Public Health, National Practice, Prevention, Preparedness, Policies and interregional coordination, 6) Research and development, 7) Psychological support, 8) Post-incident recovery and rehabilitation, 9) Interagency cooperation and coordination, and 10) Legal and ethical considerations.

## 2.4 | Statistical Analysis

Data analysis was performed using R<sup>®</sup> programming language and accessed through R-Studio<sup>®</sup>. First, descriptive statistics were determined for each agreement metric to provide an overview of the agreement levels and their precision, including counts, median, means, standard deviation, IQR, agreement percentages, the agreement metrics coefficients (Kendall's W, ICC and Kappa) and their agreement classifications (Supporting Information: Appendix 1). Second, Bland-Altman plots were designed to visualise and quantify the relationships between pairs of metrics (Lyon et al. 2023). Bland-Altman plots assess the agreement between each two of the utilised metrics to enable the identification of differences. Third, the Friedman test was employed to determine if there were statistically significant differences among the four agreement measures (Agreement percentage, Kendall's W, ICC and Kappa). The Friedman is a non-parametric test that can detect differences across multiple samples without assuming normality, which allows comparing the employed agreement metrics across different questions (Diaz-Escobar et al. 2021). Fourth, Kruskal-Wallis and Dunn's post-hoc (with Bonferroni correction) tests were deployed to analyse the agreement metrics across the general metrics of the PAT (Hayes et al. 2022). Finally, supervised machine learning (SML) techniques were used to analyse and compare the performance of different agreement metrics in predicting the agreement percentage categories (“Fair”, “Moderate”, “Good”, and “Excellent”). Removing the “Poor” category was necessary to ensure model stability. The exclusion of the “Poor” category SML models was driven by its minimal representation in the data set, which posed challenges related to model stability and

performance. Including this category could have resulted in highly imbalanced class distributions, leading to unreliable predictions and increased noise during training, as observed in studies dealing with imbalanced datasets (Abdul Manap et al. 2024; Salmi et al. 2024). Other techniques, such as over-sampling or synthetic minority oversampling technique (SMOTE), were considered but not applied due to the risk of introducing artificial bias in our expert-based data set and potentially overfitting the models to synthetic data points that may not accurately represent real-world expert opinions in CBRN PAT. Preliminary tests were conducted to evaluate the impact of retaining the “Poor” category, which revealed significant reductions in model accuracy and increased variability across predictions. Five SMLs were deployed: Decision Tree (DT), Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Naive Bayes (NB), and Gradient Boosting (GB). These algorithms were trained to predict agreement levels (Fair, Moderate, Good, and Excellent) based on three agreement metrics tested in this study. Confusion matrices were utilised to differentiate between each agreement level's correct and incorrect predictions (Farhat et al. 2024). The prediction accuracy of each SML algorithm was calculated.

Additionally, a feature importance plot was created to measure the contribution of each agreement metric to the model's decisions (Vellido 2020).

## 3 | Results

Forty experts in disaster medicine participated in this study, who evaluated 133 detailed metrics/items distributed across ten general metrics using a 5-point Likert scale. The overall mean score was  $4.91 \pm 0.71$ , with a median IQR of 0.98. The average agreement percentage was 89.21%, with Kendall's W, ICC, and Kappa coefficients of 0.50, 0.48, and 0.51, respectively. The summary statistics of the PAT general and specific metrics are presented in Table 1. The Bland-Altman plots (Figure 1) demonstrated minimal differences between pairs of agreement metrics (mean differences: Kappa vs ICC:  $-0.01$ ; ICC vs Kendall's W:  $-0.01$ ; Kappa vs Kendall's W:  $0.02$ ). Minor variability was observed in individual measurements, with 95% limits of agreement ranging from  $-0.88$  to  $0.87$  for Kappa vs ICC,  $-0.82$  to  $0.79$  for ICC vs Kendall's W, and  $-0.80$  to  $0.77$  for Kappa vs Kendall's W (Figure 2). Furthermore, the Friedman test in Table 2 revealed significant differences between the three agreement metrics employed in this study ( $\chi^2 = 302.71$ ,  $df = 3$ ,  $p < 0.05$ ). Besides that, the Kruskal-Wallis test identified that the Kappa coefficient demonstrated the most significant variation between the general metrics ( $p = 0.03$ , 95% CI:  $-0.10$  to  $0.24$ ), compared to Kendall's W and ICC ( $p = 0.12$ , 95% CI:  $-0.10$  to  $0.24$ ;  $p = 0.38$ , 95% CI:  $-0.10$  to  $0.24$ , respectively). Dunn's post-hoc test reinforced these findings. Kappa displayed more varied results and significant differences between the general metrics, whereas Kendall's W and ICC showed minimal to no significant differences in pairwise comparisons (Table 2).

The SML analysis (Table 3 and Figure 2) revealed varying model performance. The KNN algorithm demonstrated the highest overall accuracy at 43.43% (95% CI: 33.50% - 53.77%), followed by the DT model with an accuracy of 42.42% (95% CI:

**TABLE 1** | Summary statistics of the general and specific metrics of the CBRN preparedness assessment tool.

<b>General metrics</b>	<b>Specific metrics</b>	<b>Mean</b>	<b>SD</b>	<b>Median</b>	<b>IQR</b>	<b>Agreement %</b>	<b>Kendall's W</b>	<b>ICC</b>	<b>Kappa</b>
I. Medical protocols and logistics	1)At minimum, the following therapies and antidotes should be available in the MENA region due to the prevalent CBRN risks	13.07	0.46	12.83	5.17	92.92	0.31	0.42	0.82
	2) Healthcare facilities located within risk areas or close to it, should ensure immediate isolation capabilities availability:	4.58	0.41	5	1	90	0.73	0.59	0.97
	3)In the context of the MENA region, in case of CBRN emergencies:	4.69	0.74	5	0.5	93.75	0.57	0.17	0.06
II. Infrastructure Readiness for CBRN Incidents in the MENA Region	1) The following elements should be included as key indicators for health sector readiness for CBRN	4.88	0.4	5	0	99	0.46	0.35	0.16
	2) When dealing with CBRN incidents, the bed capacity and isolation unit availability are good metrics to classify hospitals in the MENA region.	4.48	0.43	5	1	86.88	0.61	0.55	0.51
	3) Each isolation room should be equipped with the following	4.73	0.36	5	0.21	94.17	0.31	0.58	0.3
III. Decontamination capabilities	1)For contaminated VIP patients, such as presidents and ministers, it is preferable to treat them separately for security reasons.	4.78	0.45	5	0	95	0.15	0.97	0.81
	2) Decontamination kits for VIPs (Such as presidents and ministers) must be provided by:	4.08	0.66	4.5	1.5	73.75	0.64	0.24	0.36
	3) Shower units should be available in the emergency department of different levels of healthcare facilities, ready to be used for patient decontamination when needed and maintained by periodic special disinfection. This statement is valid for the following hospital levels	4.52	0.69	5	0.8	87.5	0.56	0.43	0.26
	4)The shower facilities should be equipped with self-decontamination kits that include the following	4.68	0.93	5	0.25	92	0.56	0.54	0.52
	5) The definition of vulnerable contaminated patients should be expanded to include	4.67	0.75	5	0.29	92.5	0.52	0.49	0.45
	6)A range of appropriate guides to decontamination should be developed in	3.56	0.68	4.25	3.5	57.5	0.82	0.67	0.67

(Continues)

TABLE 1 | (Continued)

General metrics	Specific metrics	Mean	SD	Median	IQR	Agreement %	Kendall's W	ICC	Kappa
	consultation with representatives of the vulnerable groups. Then should be clearly visualised and easy to access when needed.								
	7) Pre-hospital Shower units should be provided by	3.51	0.65	3.83	2.67	59.17	0.42	0.48	0.28
	8) Level C personal protective equipment is sufficiently safe for a healthcare provider in a healthcare facility undertaking the decontamination of a patient following the removal of contaminated clothing and other sources	4.45	0.55	5	1	85	0.21	0.78	0.55
IV. Post-incident recovery and rehabilitation	1) There should be a system/plan for long-term medical follow-up of CBRN incident victims.	4.78	0.62	5	0	100	0.26	0.57	0.68
	2) Rehabilitation services for psychological recovery should be made accessible to CBRN victims.	4.75	0.42	5	0.25	100	0.67	0.38	0.23
	3) Procedures and resource databases for safely and efficiently restoring areas affected by CBRN incidents should be established.	4.8	1.83	5	0	100	0.72	0.41	0.92
V. Specialised Human Resources Capabilities	1) Training about Recognition and medical management of CBRN agents and Decontamination (With emphasis on hair and under the nails) should be provided to:	4.61	1.8	4.88	0.56	93.75	0.74	0.32	0.54
	3)Considering the management of a CBRN incident in a healthcare facility, for each department, what percentage of the staff per shift should have specific training in CBRN response:	4.61	0.42	4.75	0.5	91.25	0.58	0.26	0.5
VI. Public Health, National Practice, Prevention, Preparedness, Policies and interregional Coordination	1) Local contingency plans should be prepared according to the most common chemical threats and should include.	4.55	0.49	4.93	0.71	86.79	0.36	0.47	0.53
	10) Do you believe online platforms dedicated to CBRN preparedness would be beneficial for hosting webinars?	4.28	1.31	4.5	1	82.5	0.78	0.86	0.78

(Continues)

TABLE 1 | (Continued)

General metrics	Specific metrics	Mean	SD	Median	IQR	Agreement %	Kendall's W	ICC	Kappa
	11) Do you support the idea that online platforms for CBRN preparedness should offer training and best practice guidelines in CBRN preparedness?	3.83	0.63	4	2.25	65	0.7	0.73	0.45
	12) The database should include the following (local and regional levels)	4.6	0.52	4.94	0.78	92.69	0.58	0.56	0.48
	13) Local health authorities should conduct regular public health education and awareness campaigns on CBRN incidents.	4.6	0.44	5	1	95	0.79	0.42	0.86
	14) Schools and other educational institutions should include basic information about CBRN safety in their curriculum.	4.78	0.92	5	0	100	0.17	0.88	0.72
	15) There should be a protocol for early detection and surveillance of potential CBRN threats.	4.65	1.08	5	0.25	92.5	0.41	0.59	0.77
	16) Local health authorities should have a clear and effective communication strategy for potential CBRN threats.	4.6	0.62	5	1	95	0.54	0.72	0.97
	17) Community-based preparedness programs for CBRN incidents should be implemented.	4.72	0.46	5	0	95	0.55	0.32	0.39
	2) Hospitals should arrange periodic meetings to review the contingency plan at least.	4.77	0.44	5	0.25	96.88	0.52	0.36	0.49
	3) The private healthcare sector should participate in the local contingency plan preparation.	4.2	0.45	5	1.25	75	0.61	0.14	0.45
	4) Civil defence, Ministry of Interior should be included in the periodic hospital contingency plans.	4.47	0.62	5	1	87.5	0.66	1	0.14
	5) In areas with high industrial activities, Industrial facility/Factory leaders must be included in an annual planning evaluation meeting.	3.17	0.54	2.5	3	45	0.63	0.17	0.88
	6) How frequently should a national database	4.17	0.83	4.33	1.33	74.17	0.34	0.57	0.67
	7) The creation of liaison officer roles within the hospital, local, national and inter-MENA	4.07	0.7	4.25	1.56	74.38	0.19	0.3	0.47

(Continues)

TABLE 1 | (Continued)

General metrics	Specific metrics	Mean	SD	Median	IQR	Agreement %	Kendall's W	ICC	Kappa
	is helpful and should accomplish the following conditions:								
	8)The position of the liaison officer is mainly to liaise between the healthcare representatives in different sectors which should be under	4.73	1.06	5	0.25	97.5	0.57	0.59	0.55
	9) Regional knowledge hubs dedicated to CBRN preparedness are necessary?	4.28	0.36	5	1	82.5	0.19	0.27	0.45
VII. Research and Development	1) There should be sufficient resources such as funding, human capital etc to ensure continual research and development to improve CBRN response measures.	4.65	0.68	5	1	92.5	0.82	0.13	0.98
	2) Collaboration with international research organisations is needed to share knowledge and best practices in CBRN preparedness and response.	4.78	0.48	5	0	100	0.33	0.26	0.76
	3) Adopting and implementing the latest CBRN detection, decontamination, and treatment technologies should be considered.	3.5	0.52	5	4	60	0.2	0.23	0.32
VIII. Psychological Support	1) AA psychological support plan should be available for individuals directly affected by a CBRN incident such as responders, public etc	3.52	0.44	5	4	62.5	0.07	0.35	0.9
	2) Mental health professionals should be trained to handle the psychological effects of CBRN incidents on victims and their families.	4.72	0.81	5	1	100	0.12	0.81	0.82
IX. Interagency Cooperation and Coordination	1) A clear interagency cooperation and coordination framework should be created before and implemented during a CBRN incident	4.75	1.18	5	0	90	0.75	0.82	0.74
	2) Regular interagency training and simulation exercises for CBRN incidents should be conducted.	4.85	1.55	5	0	100	0.95	0.69	0.28
	3) A unified command structure should be implemented during a CBRN incident to ensure effective coordination of response efforts.	4.75	0.53	5	0.25	100	0.97	0.35	0.2

(Continues)

TABLE 1 | (Continued)

General metrics	Specific metrics	Mean	SD	Median	IQR	Agreement %	Kendall's W	ICC	Kappa
	4) Mechanisms for sharing information and resources among agencies during a CBRN incident should be established.	4.75	0.41	5	0	95	0.07	0.21	0.71
X. Legal and Ethical Considerations	1) Legal measures should be in place to enforce the implementation of safety standards and protocols in facilities dealing with CBRN materials.	4.8	0.93	5	0	97.5	0.17	0.87	0.46
	2) Ethical guidelines for handling CBRN incidents should be established, especially regarding access to treatment and rehabilitation and victims' rights.	4.7	1.09	5	1	97.5	0	0.34	0.49
	3) Privacy, data protection, and informed consent issues should be addressed in the CBRN incident response.	4.88	0.51	5	0	95	0.41	0.1	0.87

32.55% - 52.77%). The DT exhibited the highest sensitivity (0.82) for the “Excellent” class of the agreement categories. At the same time, KNN showed a more balanced performance across all classes, with the highest sensitivity (0.67) for the “Moderate” class. The SVM demonstrated a moderate sensitivity for the “Good” and “Excellent” classes (0.45 and 0.52, respectively). The GB showed high sensitivity (0.94) for the “Excellent” class but performed poorly in classifying other agreement classes. The NB demonstrated the weakest performance with an accuracy of 9.09% (95% CI: 4.24% - 16.56%).

Feature importance analysis revealed that the DT model assigned the highest importance to Kappa (168.32), followed by Kendall's W (159.10) and ICC (154.38). SVM and NB models showed negative importance for some features, indicating potential issues in their feature evaluation processes. Overall, KNN and DT algorithms performed better than the others, with the SML analysis identifying Kappa followed by Kendall's-W as the most crucial features in the DT model.

## 4 | Discussion

### 4.1 | Key Findings and Metric Sensitivities

Our study employed three primary agreement metrics: Kendall's W, ICC, and Kappa. The Friedman test ( $\chi^2 = 302.71$ ,  $df = 3$ ,  $p < 0.05$ ) indicated significant differences among these metrics, suggesting they capture distinct aspects of the agreement. The Kruskal-Wallis test helped explore these differences further, Kappa demonstrated the most significant variations across categories ( $p = 0.03$ ), unlike Kendall's W ( $p = 0.12$ ) and ICC ( $p = 0.38$ ). Kappa was demonstrated as the most sensitive metric in detecting differences across the CBRN PAT themes. Recent studies identified that Kappa's greater sensitivity may stem from its ability to account for chance agreements in complex multi-categorical assessments, common in CBRN preparedness, where expert opinions might be influenced by contextual factors such as differing local infrastructure (Hinz et al. 2021; Smith et al. 2021). The sensitivity of Kappa may be attributed to its ability to account for chance agreement, which is particularly relevant in expert assessments on complex preparedness-related issues.

While Kappa was identified as the most sensitive metric in detecting variations across CBRN PAT themes, its reliance on chance agreement adjustments may not always be ideal. For example, in situations where ordinal data is predominant or when the focus is on rank ordering rather than categorical agreement, Kendall's W may be more appropriate due to its ability to account for the order of ratings (Ruan et al. 2022). Similarly, ICC excels in scenarios requiring assessments of continuous data or when measuring consistency across multiple raters. These metrics provide stability and reliability in contexts where nuanced differences are less critical, but overall agreement is paramount, such as when evaluating standardised protocols or infrastructure readiness (Gottlieb et al. 2021). Therefore, while Kappa has demonstrated advantages in multi-categorical assessments, its application should be complemented by other metrics depending on the specific context and nature of the data.



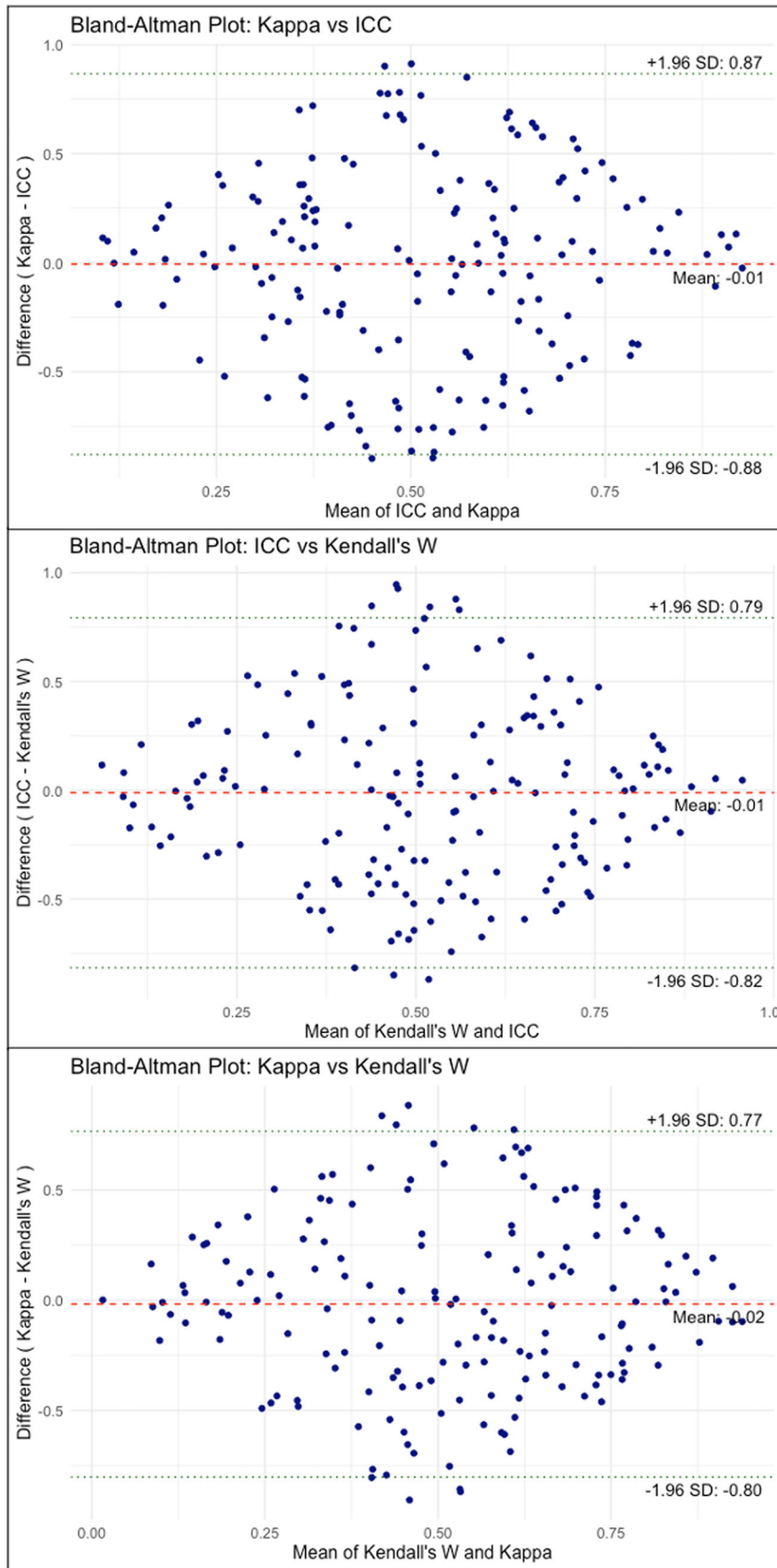


FIGURE 1 | Bland Altman plots of the agreement metrics.

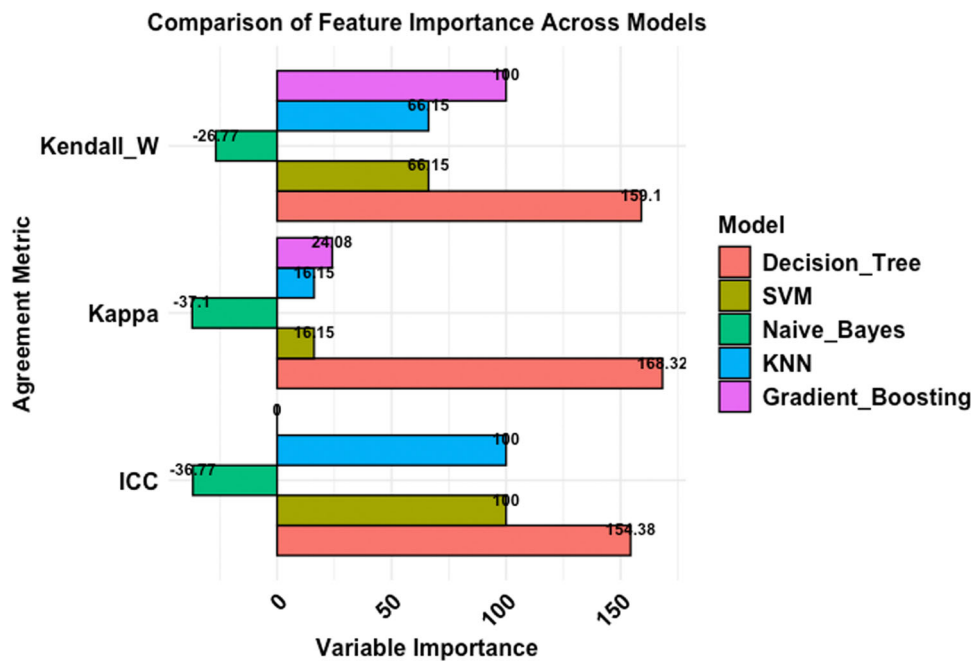


FIGURE 2 | Comparison of the agreement metrics using supervised machine learning feature importance.

#### 4.2 | Stability and Limitations of Metrics

Conversely, Kendall’s W and ICC demonstrated less variation across themes, which might also be a strength, indicating a more stable measure across diverse CBRN PAT themes. However, it could also suggest a lack of discriminatory power in detecting subtle differences between categories, as demonstrated in studies about healthcare preparedness assessments (Gupta and Federman 2020; Munasinghe et al. 2023). Further, the post-hoc Dunn’s test revealed significant differences in agreement levels across various CBRN PAT themes, such as specialised human resources capabilities, psychological support, medical protocols and logistics, infrastructure readiness, and public health policies. These variations highlight the complexity of CBRN preparedness and the challenges in achieving a perfect consensus across these complex themes. For instance, significant differences in consensus were observed among experts between “Medical Protocols and Logistics” and “Infrastructure Readiness for CBRN Incidents” (Kappa analysis,  $p = 0.08$ ), highlighting the disparity in expert opinions regarding operational versus infrastructural aspects of preparedness. The infrastructural capabilities might not always adequately fulfil the operational needs and may not always be sufficient to support the implementation of an operational plan effectively. This might be due to the heterogeneous infrastructural capabilities across MENA countries and the lack of unified operational response guidelines that consider these variations. Addressing this variation necessitates the development of standardised approaches that account for local and regional differences in resources, expertise, and infrastructure. This can be accomplished by promoting collaborative research and targeted improvement initiatives in regional disaster preparedness. Such efforts should focus on the themes where a lack of consensus was identified within the general metrics of the PAT in this study and as emphasised by similar studies (Dinar et al. 2023).

#### 4.3 | Contextual Challenges in Expert Consensus

Overall, the variant performance of agreement metrics in disaster medicine preparedness studies reveals profound challenges beyond mathematical precision, revealing complexities in expert knowledge aggregation. The metrics employed in this study illustrate difficulties of consensus formation due to contextual variations, regional infrastructure disparities, and evolving threat nature that influence expert perspectives. The differential sensitivity of consensus metrics, evident in this study and other previous studies, exposes challenges in developing unified emergency response strategies (Grodman et al. 2023; Hung et al. 2022). Factors such as professional diversity, variance of capabilities between countries, and geopolitical constraints could contribute to the variability in expert consensus. Nevertheless, while this variability can present potential risks—including fragmented perspectives and inconsistent preparedness protocols—it offers opportunities for methodological innovation, such as implementing context-aware algorithmic approaches. These algorithmic approaches represent an intelligent computational example that dynamically adapts technological responses by synthesising multiple contextual variables beyond traditional linear processing, analysing environmental, user, and systemic data to generate personalised interactions (Gulati and Raman 2024; Zon et al. 2023). In disaster medicine, these approaches can integrate geospatial data, historical incident reports, regional infrastructure capabilities, and real-time expert input to generate dynamically adjusted emergency response strategies, which allow for predicting potential scenarios and recommend adaptive interventions that traditional static models cannot achieve.

#### 4.4 | Machine Learning Benefits and Outcomes

Additionally, applying SML algorithms provided deeper perspectives on the performance of agreement metrics. The KNN

**TABLE 2** | Comparison between the Delphi agreement metrics using statistical analysis.

	<b>Statistic</b>	<b>Value</b>	
<b>I. Friedman test results</b>			
Friedman $\chi^2$	$\chi^2$	302.71	
df	Degrees of freedom	3	
	<i>p</i> -value	< 0.05	
<b>Agreement metric</b>	<b><i>p</i>-value</b>	<b>CI Lower</b>	<b>CI Upper</b>
<b>II. Kruskal-Wallis Test Results</b>			
Kendall's W	0.12	-0.10	0.24
ICC	0.38	-0.10	0.24
Kappa	0.03	-0.10	0.24
<b>Comparison</b>	<b>Z-value</b>	<b>p. unadj</b>	<b>Adjusted p-value</b>
<b>III. Post-hoc Dunn's Test Results</b>			
1) Kendall's W			
II. Medical Protocols and Logistics - III. Infrastructure Readiness for CBRN Incidents in the MENA Region	-0.14	0.89	1.00
II. Medical Protocols and Logistics - IV. Decontamination capabilities	-1.02	0.31	1.00
III. Infrastructure Readiness for CBRN Incidents in the MENA Region - IV. Decontamination capabilities	-1.03	0.30	1.00
II. Medical Protocols and Logistics - IX. Post-Incident Recovery and Rehabilitation	-0.83	0.41	1.00
III. Infrastructure Readiness for CBRN Incidents in the MENA Region - IX. Post-Incident Recovery and Rehabilitation	-0.78	0.44	1.00
IV. Decontamination capabilities - IX. Post-Incident Recovery and Rehabilitation	-0.26	0.80	1.00
II. Medical Protocols and Logistics - V. Specialised Human Resources Capabilities	-1.73	0.08	1.00
III. Infrastructure Readiness for CBRN Incidents in the MENA Region - V. Specialised Human Resources Capabilities	-1.79	0.07	1.00
IV. Decontamination capabilities - V. Specialised Human Resources Capabilities	-1.11	0.27	1.00
IX. Post-Incident Recovery and Rehabilitation - V. Specialised Human Resources Capabilities	-0.43	0.67	1.00
II. Medical Protocols and Logistics - VI. Public Health, National Practice, Prevention, Preparedness, Policies and interregional Coordination	-0.92	0.36	1.00
III. Infrastructure Readiness for CBRN Incidents in the MENA Region - VI. Public Health, National Practice, Prevention, Preparedness, Policies and interregional Coordination	-0.93	0.35	1.00
IV. Decontamination capabilities - VI. Public Health, National Practice, Prevention, Preparedness, Policies and interregional Coordination	0.28	0.78	1.00
IX. Post-Incident Recovery and Rehabilitation - VI. Public Health, National Practice, Prevention, Preparedness, Policies and interregional Coordination	0.38	0.70	1.00
V. Specialised Human Resources Capabilities - VI. Public Health, National Practice, Prevention, Preparedness, Policies and interregional Coordination	1.36	0.173	1.00

(Continues)

TABLE 2 | (Continued)

Comparison	Z-value	p. unadj	Adjusted p-value
II. Medical Protocols and Logistics - VII. Research and Development	-0.26	0.80	1.00
III. Infrastructure Readiness for CBRN Incidents in the MENA Region - VII. Research and Development	-0.18	0.856	1.00
IV. Decontamination capabilities - VII. Research and Development	0.36	0.72	1.00
IX. Post-Incident Recovery and Rehabilitation - VII. Research and Development	0.47	0.65	1.00
V. Specialised Human Resources Capabilities - VII. Research and Development	0.99	0.32	1.00
VI. Public Health, National Practice, Prevention, Preparedness, Policies and interregional Coordination - VII. Research and Development	0.26	0.79	1.00
II. Medical Protocols and Logistics - VIII. Psychological Support	1.50	0.13	1.00
III. Infrastructure Readiness for CBRN Incidents in the MENA Region - VIII. Psychological Support	1.64	0.10	1.00
IV. Decontamination capabilities - VIII. Psychological Support	2.14	0.03	1.00
IX. Post-Incident Recovery and Rehabilitation - VIII. Psychological Support	1.89	0.06	1.00
V. Specialised Human Resources Capabilities - VIII. Psychological Support	2.55	0.01	0.49
VI. Public Health, National Practice, Prevention, Preparedness, Policies and interregional Coordination - VIII. Psychological Support	2.09	0.04	1.00
VII. Research and Development - VIII. Psychological Support	1.47	0.14	1.00
II. Medical Protocols and Logistics - X. Interagency Cooperation and Coordination	-1.61	0.11	1.00
III. Infrastructure Readiness for CBRN Incidents in the MENA Region - X. Interagency Cooperation and Coordination	-1.61	0.11	1.00
IV. Decontamination capabilities - X. Interagency Cooperation and Coordination	-1.07	0.29	1.00
IX. Post-Incident Recovery and Rehabilitation - X. Interagency Cooperation and Coordination	-0.54	0.59	1.00
V. Specialised Human Resources Capabilities - X. Interagency Cooperation and Coordination	-0.21	0.84	1.00
VI. Public Health, National Practice, Prevention, Preparedness, Policies and interregional Coordination - X. Interagency Cooperation and Coordination	-1.24	0.22	1.00
VII. Research and Development - X. Interagency Cooperation and Coordination	-1.04	0.30	1.00
VIII. Psychological Support - X. Interagency Cooperation and Coordination	-2.47	0.01	0.60
II. Medical Protocols and Logistics - XI. Legal and Ethical Considerations	1.14	0.26	1.00
III. Infrastructure Readiness for CBRN Incidents in the MENA Region - XI. Legal and Ethical Considerations	1.30	0.19	1.00
IV. Decontamination capabilities - XI. Legal and Ethical Considerations	1.90	0.06	1.00
IX. Post-Incident Recovery and Rehabilitation - XI. Legal and Ethical Considerations	1.61	0.11	1.00
V. Specialised Human Resources Capabilities - XI. Legal and Ethical Considerations	2.37	0.02	0.80

(Continues)

TABLE 2 | (Continued)

Comparison	Z-value	p. unadj	Adjusted p-value
VI. Public Health, National Practice, Prevention, Preparedness, Policies and interregional Coordination - XI. Legal and Ethical Considerations	1.84	0.07	1.00
VII. Research and Development - XI. Legal and Ethical Considerations	1.14	0.25	1.00
VIII. Psychological Support - XI. Legal and Ethical Considerations	-0.45	0.65	1.00
X. Interagency Cooperation and Coordination - XI. Legal and Ethical Considerations	2.26	0.02	1.00
2) ICC			
II. Medical Protocols and Logistics - III. Infrastructure Readiness for CBRN Incidents in the MENA Region	-0.97	0.33	1.00
II. Medical Protocols and Logistics - IV. Decontamination capabilities	-1.10	0.27	1.00
III. Infrastructure Readiness for CBRN Incidents in the MENA Region - IV. Decontamination capabilities	-0.06	0.95	1.00
II. Medical Protocols and Logistics - IX. Post-Incident Recovery and Rehabilitation	-0.53	0.60	1.00
III. Infrastructure Readiness for CBRN Incidents in the MENA Region - IX. Post-Incident Recovery and Rehabilitation	0.09	0.93	1.00
IV. Decontamination capabilities - IX. Post-Incident Recovery and Rehabilitation	0.12	0.90	1.00
II. Medical Protocols and Logistics - V. Specialised Human Resources Capabilities	0.70	0.49	1.00
III. Infrastructure Readiness for CBRN Incidents in the MENA Region - V. Specialised Human Resources Capabilities	1.70	0.09	1.00
IV. Decontamination capabilities - V. Specialised Human Resources Capabilities	1.89	0.06	1.00
IX. Post-Incident Recovery and Rehabilitation - V. Specialised Human Resources Capabilities	1.02	0.31	1.00
II. Medical Protocols and Logistics - VI. Public Health, National Practice, Prevention, Preparedness, Policies and interregional Coordination	-1.37	0.17	1.00
III. Infrastructure Readiness for CBRN Incidents in the MENA Region - VI. Public Health, National Practice, Prevention, Preparedness, Policies and interregional Coordination	-0.28	0.79	1.00
IV. Decontamination capabilities - VI. Public Health, National Practice, Prevention, Preparedness, Policies and interregional Coordination	-0.26	0.7	1.00
IX. Post-Incident Recovery and Rehabilitation - VI. Public Health, National Practice, Prevention, Preparedness, Policies and interregional Coordination	-0.23	0.82	1.00
V. Specialised Human Resources Capabilities - VI. Public Health, National Practice, Prevention, Preparedness, Policies and interregional Coordination	-2.20	0.03	1.00
II. Medical Protocols and Logistics - VII. Research and Development	0.96	0.33	1.00
III. Infrastructure Readiness for CBRN Incidents in the MENA Region - VII. Research and Development	1.66	0.10	1.00
IV. Decontamination capabilities - VII. Research and Development	1.75	0.08	1.00
IX. Post-Incident Recovery and Rehabilitation - VII. Research and Development	1.22	0.22	1.00
V. Specialised Human Resources Capabilities - VII. Research and Development	0.45	0.65	1.00

(Continues)

TABLE 2 | (Continued)

Comparison	Z-value	p. unadj	Adjusted p-value
VI. Public Health, National Practice, Prevention, Preparedness, Policies and interregional Coordination - VII. Research and Development	1.91	0.06	1.00
II. Medical Protocols and Logistics - VIII. Psychological Support	-0.93	0.35	1.00
III. Infrastructure Readiness for CBRN Incidents in the MENA Region - VIII. Psychological Support	-0.43	0.67	1.00
IV. Decontamination capabilities - VIII. Psychological Support	-0.41	0.69	1.00
IX. Post-Incident Recovery and Rehabilitation - VIII. Psychological Support	-0.41	0.69	1.00
V. Specialised Human Resources Capabilities - VIII. Psychological Support	-1.35	0.18	1.00
VI. Public Health, National Practice, Prevention, Preparedness, Policies and interregional Coordination - VIII. Psychological Support	-0.33	0.74	1.00
VII. Research and Development - VIII. Psychological Support	-1.50	0.13	1.00
II. Medical Protocols and Logistics - X. Interagency Cooperation and Coordination	-0.77	0.44	1.00
III. Infrastructure Readiness for CBRN Incidents in the MENA Region - X. Interagency Cooperation and Coordination	-0.10	0.92	1.00
IV. Decontamination capabilities - X. Interagency Cooperation and Coordination	-0.07	0.95	1.00
IX. Post-Incident Recovery and Rehabilitation - X. Interagency Cooperation and Coordination	-0.14	0.89	1.00
V. Specialised Human Resources Capabilities - X. Interagency Cooperation and Coordination	-1.31	0.19	1.00
VI. Public Health, National Practice, Prevention, Preparedness, Policies and interregional Coordination - X. Interagency Cooperation and Coordination	0.05	0.96	1.00
VII. Research and Development - X. Interagency Cooperation and Coordination	-1.45	0.15	1.00
VIII. Psychological Support - X. Interagency Cooperation and Coordination	0.31	0.76	1.00
II. Medical Protocols and Logistics - XI. Legal and Ethical Considerations	-0.28	0.78	1.00
III. Infrastructure Readiness for CBRN Incidents in the MENA Region - XI. Legal and Ethical Considerations	0.35	0.73	1.00
IV. Decontamination capabilities - XI. Legal and Ethical Considerations	0.39	0.70	1.00
IX. Post-Incident Recovery and Rehabilitation - XI. Legal and Ethical Considerations	0.20	0.84	1.00
V. Specialised Human Resources Capabilities - XI. Legal and Ethical Considerations	-0.78	0.44	1.00
VI. Public Health, National Practice, Prevention, Preparedness, Policies and interregional Coordination - XI. Legal and Ethical Considerations	0.51	0.61	1.00
VII. Research and Development - XI. Legal and Ethical Considerations	-1.02	0.31	1.00
VIII. Psychological Support - XI. Legal and Ethical Considerations	0.59	0.55	1.00
X. Interagency Cooperation and Coordination - XI. Legal and Ethical Considerations	0.36	0.72	1.00
3) Kappa			

(Continues)

TABLE 2 | (Continued)

Comparison	Z-value	p. unadj	Adjusted p-value
II. Medical Protocols and Logistics - III. Infrastructure Readiness for CBRN Incidents in the MENA Region	3.12	< 0.005	0.08
II. Medical Protocols and Logistics - IV. Decontamination capabilities	2.20	0.03	1.00
III. Infrastructure Readiness for CBRN Incidents in the MENA Region - IV. Decontamination capabilities	-1.44	0.15	1.00
II. Medical Protocols and Logistics - IX. Post-Incident Recovery and Rehabilitation	0.32	0.75	1.00
III. Infrastructure Readiness for CBRN Incidents in the MENA Region - IX. Post-Incident Recovery and Rehabilitation	-1.75	0.08	1.00
IV. Decontamination capabilities - IX. Post-Incident Recovery and Rehabilitation	-1.05	0.30	1.00
II. Medical Protocols and Logistics - V. Specialised Human Resources Capabilities	1.15	0.25	1.00
III. Infrastructure Readiness for CBRN Incidents in the MENA Region - V. Specialised Human Resources Capabilities	-1.73	0.08	1.00
IV. Decontamination capabilities - V. Specialised Human Resources Capabilities	-0.73	0.47	1.00
IX. Post-Incident Recovery and Rehabilitation - V. Specialised Human Resources Capabilities	0.51	0.61	1.00
II. Medical Protocols and Logistics - VI. Public Health, National Practice, Prevention, Preparedness, Policies and interregional Coordination	1.40	0.16	1.00
III. Infrastructure Readiness for CBRN Incidents in the MENA Region - VI. Public Health, National Practice, Prevention, Preparedness, Policies and interregional Coordination	-2.83	< 0.005	0.21
IV. Decontamination capabilities - VI. Public Health, National Practice, Prevention, Preparedness, Policies and interregional Coordination	-1.49	0.14	1.00
IX. Post-Incident Recovery and Rehabilitation - VI. Public Health, National Practice, Prevention, Preparedness, Policies and interregional Coordination	0.49	0.63	1.00
V. Specialised Human Resources Capabilities - VI. Public Health, National Practice, Prevention, Preparedness, Policies and interregional Coordination	-0.15	0.88	1.00
II. Medical Protocols and Logistics - VII. Research and Development	0.00	0.99	1.00
III. Infrastructure Readiness for CBRN Incidents in the MENA Region - VII. Research and Development	-2.09	0.04	1.00
IV. Decontamination capabilities - VII. Research and Development	-1.40	0.16	1.00
IX. Post-Incident Recovery and Rehabilitation - VII. Research and Development	-0.26	0.79	1.00
V. Specialised Human Resources Capabilities - VII. Research and Development	-0.83	0.41	1.00
VI. Public Health, National Practice, Prevention, Preparedness, Policies and interregional Coordination - VII. Research and Development	-0.85	0.39	1.00
II. Medical Protocols and Logistics - VIII. Psychological Support	-0.93	0.35	1.00
III. Infrastructure Readiness for CBRN Incidents in the MENA Region - VIII. Psychological Support	-2.72	0.01	0.30
IV. Decontamination capabilities - VIII. Psychological Support	-2.15	0.03	1.00

(Continues)

TABLE 2 | (Continued)

Comparison	Z-value	p. unadj	Adjusted p-value
IX. Post-Incident Recovery and Rehabilitation - VIII. Psychological Support	-1.03	0.30	1.00
V. Specialised Human Resources Capabilities - VIII. Psychological Support	-1.63	0.10	1.00
VI. Public Health, National Practice, Prevention, Preparedness, Policies and interregional Coordination - VIII. Psychological Support	-1.71	0.08	1.00
VII. Research and Development - VIII. Psychological Support	-0.80	0.42	1.00
II. Medical Protocols and Logistics - X. Interagency Cooperation and Coordination	1.12	0.26	1.00
III. Infrastructure Readiness for CBRN Incidents in the MENA Region - X. Interagency Cooperation and Coordination	-1.15	0.25	1.00
IV. Decontamination capabilities - X. Interagency Cooperation and Coordination	-0.33	0.74	1.00
IX. Post-Incident Recovery and Rehabilitation - X. Interagency Cooperation and Coordination	0.60	0.55	1.00
V. Specialised Human Resources Capabilities - X. Interagency Cooperation and Coordination	0.19	0.85	1.00
VI. Public Health, National Practice, Prevention, Preparedness, Policies and interregional Coordination - X. Interagency Cooperation and Coordination	0.33	0.74	1.00
VII. Research and Development - X. Interagency Cooperation and Coordination	0.88	0.38	1.00
VIII. Psychological Support - X. Interagency Cooperation and Coordination	1.62	0.11	1.00
II. Medical Protocols and Logistics - XI. Legal and Ethical Considerations	0.35	0.72	1.00
III. Infrastructure Readiness for CBRN Incidents in the MENA Region - XI. Legal and Ethical Considerations	-1.72	0.09	1.00
IV. Decontamination capabilities - XI. Legal and Ethical Considerations	-1.02	0.31	1.00
IX. Post-Incident Recovery and Rehabilitation - XI. Legal and Ethical Considerations	0.02	0.98	1.00
V. Specialised Human Resources Capabilities - XI. Legal and Ethical Considerations	-0.48	0.63	1.00
VI. Public Health, National Practice, Prevention, Preparedness, Policies and interregional Coordination - XI. Legal and Ethical Considerations	-0.46	0.65	1.00
VII. Research and Development - XI. Legal and Ethical Considerations	0.29	0.77	1.00
VIII. Psychological Support - XI. Legal and Ethical Considerations	1.05	0.29	1.00
X. Interagency Cooperation and Coordination - XI. Legal and Ethical Considerations	-0.58	0.56	1.00

and DT algorithms demonstrated the highest overall accuracy (43.43% and 42.42%, respectively). However, this low accuracy across all models suggests limitations (and cautions) in using these metrics as unique predictors of agreement levels. These limitations can be overcome, for example, through continuous assessments conducted over time and tracking the changes in agreement levels could provide insights into evolving consensus or divergence. Furthermore, the feature importance analysis from the DT model assigned the highest importance to Kappa (168.32), followed by Kendall's W (159.10) and ICC (154.38), reaffirming the statistical

findings in this study and further supporting the notion that Kappa may be the most discriminative metric in identifying experts consensus in disaster preparedness studies.

#### 4.5 | Consensus Analysis Using Machine Learning: Opportunities for Innovation

Moreover, integrating ML algorithms in our analysis represents a transformative approach to consensus assessment in the



**TABLE 3** | Comparison between the Delphi agreement analysis metrics using supervised machine learning.

	Reference				
	Poor	Fair	Moderate	Good	Excellent
<b>I. Decision tree</b>					
Prediction					
Poor	0	0	0	0	0
Fair	0	0	0	0	0
Moderate	0	0	0	0	2
Good	0	0	11	15	4
Excellent	0	0	22	18	27
Accuracy: 0.42; 95% CI: (0.33, 0.53); No Information Rate: 0.33; p [Acc > NIR]: 0.04; Kappa: 0.14					
Statistics by class					
Sensitivity	NA	NA	0	0.45	0.82
Specificity	1	1	0.97	0.77	0.39
Pos Pred Value	NA	NA	0	0.50	0.40
Neg Pred Value	NA	NA	0.66	0.74	0.81
Prevalence	0.33	0.33	0.33		
Detection Rate	0	0	0	0.15	0.27
Detection Prevalence	0	0	0.02	0.30	0.68
Balanced Accuracy	NA	NA	0.48	0.61	0.61
<b>II. Support vector machine</b>					
Prediction					
Poor	0	0	0	0	0
Fair	0	0	0	0	4
Moderate	0	0	0	0	3
Good	0	0	33	15	9
Excellent	0	0	0	18	17
Accuracy: 0.32; 95% CI: (0.23, 0.42); No Information Rate: 0.33;P-Value [Acc > NIR]: 0.62; Kappa: 0.005					
Statistics by class					
Sensitivity	NA	NA	0	0.45	0.52
Specificity	1	0.96	0.95	0.36	0.73
Pos Pred Value	NA	NA	0	0.26	0.49
Neg Pred Value	NA	NA	0.66	0.57	0.75
Prevalence	0	0	0.33	0.33	0.33
Detection Rate	0	0	0	0.15	0.17
Detection Prevalence	0	0.04	0.03	0.58	0.35
Balanced Accuracy	NA	NA	0.48	0.41	0.62
<b>III. K-Nearest neighbouring</b>					
Prediction					
Poor	0	0	0	0	0
Fair	0	0	0	0	6
Moderate	0	0	22	8	5
Good	0	0	11	10	11
Excellent	0	0	0	15	11
Accuracy: 0.43; 95% CI: (0.34, 0.54); No Information Rate: 0.33;P-Value [Acc > NIR]: 0.02; Kappa: 0.18					

(Continues)

TABLE 3 | (Continued)

	Reference				
	Poor	Fair	Moderate	Good	Excellent
Sensitivity	NA	NA	0.67	0.30	0.33
Specificity	1	0.94	0.80	0.67	0.77
Pos Pred Value	NA	NA	0.63	0.31	0.42
Neg Pred Value	NA	NA	0.83	0.66	0.70
Prevalence	0	0	0.33	0.33	0.33
Detection Rate	0	0	0.22	0.10	0.11
Detection Prevalence	0	0.06	0.35	0.32	0.26
Balanced Accuracy	NA	NA	0.73	0.48	0.55
IV. Naïve Bayes					
Prediction					
Poor	0	0	0	0	0
Fair	0	0	22	15	12
Moderate	0	0	0	8	7
Good	0	0	11	0	5
Excellent	0	0	0	10	9
Accuracy: 0.09; 95% CI: (0.04, 0.17); No Information Rate: 0.33; p-Value [Acc > NIR]: 1; Kappa: -0.09					
Sensitivity	NA	NA	0	0	0.27
Specificity	1	0.51	0.77	0.76	0.85
Pos Pred Value	NA	NA	0	0	0.47
Neg Pred Value	NA	NA	0.61	0.60	0.70
Prevalence	0	0	0.33	0.33	0.33
Detection Rate	0	0	0	0	0.09
Detection Prevalence	0	0.49	0.15	0.16	0.19
Balanced Accuracy	NA	NA	0.39	0.38	0.56
V. Gradient boosting					
Prediction					
Poor	0	0	0	0	0
Fair	0	0	0	0	0
Moderate	0	0	0	0	1
Good	0	0	11	0	1
Excellent	0	0	22	33	31
Accuracy: 0.31; 95% CI: (0.22, 0.41); No Information Rate: 0.33; p-Value [Acc > NIR]: 0.70; Kappa: -0.03					
Sensitivity	NA	NA	0	0	0.94
Specificity	1	1	0.98	0.82	0.17
Pos Pred Value	NA	NA	0	0	0.36
Neg Pred Value	NA	NA	0.66	0.62	0.85
Prevalence	0	0	0.33	0.33	0.33
Detection Rate	0	0	0	0	0.31
Detection Prevalence	0	0	0.01	0.12	0.87
Balanced Accuracy	NA	NA	0.49	0.41	0.55

Delphi studies in disaster medicine, demonstrating how artificial intelligence can enhance our understanding of expert reflections. The ML analysis is crucial to more dynamic, data-driven disaster preparedness strategies. By leveraging advanced analytical techniques, we can more effectively identify knowledge gaps, track evolving expert perspectives, and develop more context-aware emergency response protocols that rapidly adapt to emerging CBRN threats. This approach is helpful in regions with complex geopolitical and infrastructural challenges where traditional consensus methodologies may fail to capture the required readiness level, such as in the MENA region.

## 5 | Limitations

The moderately low accuracy of the machine learning models indicates a limitation in using these metrics alone as the agreement-level predictors. Further, while the “Poor” category was excluded from the SML analysis due to the models’ stability necessity, this might have affected the consensus assessment in PAT metrics with the low agreement. Using purposive and snowball sampling may have introduced selection bias in the expert panel composition. Future Delphi studies should consider a larger sample size and employ more advanced artificial intelligence analysis techniques.

## 6 | Conclusion

In this study, both statistical analyses and machine learning approaches consistently identified Kappa as the most sensitive and discriminative metric for detecting differences in expert consensus across CBRN PAT themes. Further, the SML also suggested that the tested agreement metrics alone may not be sufficient predictors of overall consensus level due to the region’s complexity of CBRN preparedness metrics.

Therefore, it is recommended that consensus metrics be complemented with continuous assessment over time when assessing experts’ opinions about disaster preparedness. Regular evaluations using statistical tools and machine learning approaches can provide a deeper understanding of evolving consensus patterns and help identify areas requiring focused intervention. Future research should explore integrating additional contextual factors, such as regional variations in infrastructure, resources, and operational guidelines, to enhance agreement metrics’ predictive power and interpretability. For instance, incorporating geopolitical constraints, such as cross-border collaboration challenges or conflict zones, and socioeconomic disparities, like funding inequities or access to specialised training, could provide a deeper understanding of regional preparedness dynamics. Additionally, cultural and logistical factors, including community trust in authorities or supply chain vulnerabilities, should be considered when adapting strategies to local realities. Furthermore, continuous refinement of assessment techniques and ongoing evaluation and adaptation to regional specificities will strengthen CBRN preparedness and response capabilities across the MENA region.

---

### Author Contributions

**Hassan Farhat:** writing – original draft, conceptualisation, data curation, formal analysis, methodology, investigation, software. **Alan M.**

**Batt, Mariana Helou, Heejun Shin, James Laughton, Carolyn Dumbeck, Arezoo Dehghani, Fatemeh Rezaei, Nidaa Bajow, Luc Mortelmans, Walid Abougalala, and Roberto Mugavero:** writing – review and editing. **Gregory Ciotto, Mohamed Ben Dhiab, and Guillaume Alinier:** writing – review and editing, supervision.

### Acknowledgements

The authors thank all the experts who participated in this study. We thank the reviewers for their feedback that helped improve the quality of the manuscript. Hamad Medical Corporation Open Access publishing facilitated by the Qatar National Library, as part of the Wiley Qatar National Library agreement.

### Ethics Statement

This study was approved by the Ethical Committees of the Faculty of Medicine “Ibn Eljazzar” of Sousse in Tunisia and Hamad Medical Corporation’s Medical Research Center in Qatar (references CEFMS 110/2022 and MRC-01-22-258, respectively).

### Conflicts of Interest

The authors declare no conflicts of interest.

### Data Availability Statement

The data is available with the first author and can be provided upon request and pending approval.

### References

- Abdul Manap, A. S., R. Almadodi, S. Sultana, et al. 2024. “Alzheimer’s Disease: A Review on the Current Trends of the Effective Diagnosis and Therapeutics.” *Frontiers in Aging Neuroscience* 16: 1429211. <https://doi.org/10.3389/fnagi.2024.1429211>.
- Alammary, A. S. 2022. “How to Decide the Proportion of Online to Face-To-Face Components of a Blended Course? A Delphi Study.” *Sage Open* 12, no. 4: 21582440221138448. <https://doi.org/10.1177/21582440221138448>.
- Banno, M., Y. Tsujimoto, and Y. Kataoka. 2020. “The Majority of Reporting Guidelines are not Developed With the Delphi Method: A Systematic Review of Reporting Guidelines.” *Journal of Clinical Epidemiology* 124: 50–57. <https://doi.org/10.1016/j.jclinepi.2020.04.010>.
- Denham, B. E. 2016. *Categorical Statistics for Communication Research*. John Wiley & Sons.
- Diaz-Escobar, J., N. E. Ordóñez-Guillén, S. Villarreal-Reyes, et al. 2021. “Deep-Learning Based Detection of COVID-19 Using Lung Ultrasound Imagery.” *PLoS One* 16, no. 8: e0255886. <https://doi.org/10.1371/journal.pone.0255886>.
- Dinar, H. A., H. Hummadi, D. A. Alonayzan, and L. S. Alharbi. 2023. “Experts’ Opinion on Disaster Risk Reduction (Drr) Educational Strategies in Middle East/North.” *Africa (MENA) Countries: A Delphi Study (SSRN Scholarly)* 1::: 4467661. <https://doi.org/10.2139/ssrn.4467661>.
- Drumm, S., C. Bradley, and F. Moriarty. 2022. “More of an Art Than a Science? The Development, Design and Mechanics of the Delphi Technique.” *Research in Social & Administrative Pharmacy: RSAP* 18, no. 1: 2230–2236. <https://doi.org/10.1016/j.sapharm.2021.06.027>.
- Farhat, H., G. Alinier, N. Bajow, et al. 2024. “Preparedness and Response Strategies for Chemical, Biological, Radiological, and Nuclear Incidents in the Middle East and North Africa: An Artificial Intelligence-Enhanced Delphi Approach.” *Disaster Medicine and Public Health Preparedness* 18: e244. <https://doi.org/10.1017/dmp.2024.160>.
- Farhat, H., A. Makhlof, P. Gangaram, et al. 2024. “Predictive Modelling of Transport Decisions and Resources Optimisation in Pre-Hospital

- Setting Using Machine Learning Techniques." *PLoS One* 19, no. 5: e0301472. <https://doi.org/10.1371/journal.pone.0301472>.
- Franc, J. M., K. K. C. Hung, A. Pirisi, and E. S. Weinstein. 2023. "Analysis of Delphi Study 7-point Linear Scale Data by Parametric Methods: Use of the Mean and Standard Deviation." *Methodological Innovations* 16, no. 2: 226–233. <https://doi.org/10.1177/20597991231179393>.
- Gottlieb, M., J. Jordan, J. N. Siegelman, R. Cooney, C. Stehman, and T. M. Chan. 2021. "Direct Observation Tools in Emergency Medicine: A Systematic Review of the Literature." *AEM Education and Training* 5, no. 3: e10519. <https://doi.org/10.1002/aet2.10519>.
- Gray, M. M., C. R. Butler, L. B. Webster, M. R. Tonelli, V. L. Sakata, and D. S. Diekema. 2023. "Patient Information Items Needed to Guide the Allocation of Scarce Life-Sustaining Resources: A Delphi Study of Multidisciplinary Experts." *Disaster Medicine and Public Health Preparedness* 17: e81. <https://doi.org/10.1017/dmp.2021.351>.
- Grodman, S., A. Hard, A. Hertelendy, et al. 2023. "Delphi Process Recommendations for Pediatric Disaster Medicine Training Curriculum Key Competencies." *Prehospital and Disaster Medicine* 38, no. S1: s182. <https://doi.org/10.1017/S1049023X23004715>.
- Gulati, J., and R. Raman. 2024. "A Context-Aware Emergency Assistance Chatbot Employing Recurrent Neural Networks for Personalized First Aid Guidance." *International Conference on E-Mobility, Power Control and Smart Systems (ICEMPS)* 2024: 1–5. <https://doi.org/10.1109/ICEMPS60684.2024.10559306>.
- Gupta, S., and D. G. Federman. 2020. "Hospital Preparedness for COVID-19 Pandemic: Experience From Department of Medicine at Veterans Affairs Connecticut Healthcare System." *Postgraduate Medicine* 132, no. 6: 489–494. <https://doi.org/10.1080/00325481.2020.1761668>.
- Hayes, S. C., M. White, C. R. J. Wilcox, H. S. F. White, and N. Vanicek. 2022. "Biomechanical Differences Between Able-Bodied and Spinal Cord Injured Individuals Walking in an Overground Robotic Exoskeleton." *PLoS One* 17, no. 1: e0262915. <https://doi.org/10.1371/journal.pone.0262915>.
- Hill, J., T. Ashken, S. West, et al. 2022. "Core Outcome Set for Peripheral Regional Anesthesia Research: A Systematic Review and Delphi Study." *Regional Anesthesia and Pain Medicine* 47, no. 11: 691–697. <https://doi.org/10.1136/rapm-2022-103751>.
- Hinz, M., N. Lehmann, K. Melcher, et al. 2021. "Reliability of Perceptual-Cognitive Skills in a Complex, Laboratory-Based Team-Sport Setting." *Applied Sciences* 11, no. 11), Article: 5203. <https://doi.org/10.3390/app11115203>.
- Hoinig, T., L. Rahlf, J. Wilke, et al. 2024. "Appraising the Methodological Quality of Sports Injury Video Analysis Studies: The QA-SIVAS Scale." *Sports Medicine* 54, no. 1: 203–211. <https://doi.org/10.1007/s40279-023-01907-z>.
- Hung, K. K. C., M. K. MacDermot, E. Y. Y. Chan, et al. 2022. "Health Emergency and Disaster Risk Management Workforce Development Strategies: Delphi Consensus Study." *Prehospital and Disaster Medicine* 37, no. 6: 735–748. <https://doi.org/10.1017/S1049023X22001467>.
- Keating, A., and S. Hanger-Kopp. 2020. "Practitioner Perspectives of Disaster Resilience in International Development." *International Journal of Disaster Risk Reduction* 42: e101355. <https://doi.org/10.1016/j.ijdr.2019.101355>.
- Lyon, M., C. A. Fehlmann, M. Augsburg, et al. 2023. "Evaluation of a Portable Blood Gas Analyzer for Prehospital Triage in Carbon Monoxide Poisoning: Instrument Validation Study." *JMIR Formative Research* 7, no. 1: e48057. <https://doi.org/10.2196/48057>.
- Mani, Z., V. Plummer, L. Kuhn, A. Khorram-Manesh, D. Tin, and K. Goniewicz. 2024. "Public Health Responses to CBRN Terrorism in the Middle East and North Africa." *Disaster Medicine and Public Health Preparedness* 18: e87. <https://doi.org/10.1017/dmp.2024.73>.
- Munasinghe, N. L., G. O'Reilly, and P. Cameron. 2023. "Examining the Components and Validity of Hospital Disaster Preparedness Tools." *Progress in Disaster Science* 20: 100305. <https://doi.org/10.1016/j.pdisas.2023.100305>.
- Munblit, D., T. Nicholson, A. Akrami, et al. 2022. "A Core Outcome Set for post-COVID-19 Condition in Adults for Use In Clinical Practice and Research: An International Delphi Consensus Study." *The Lancet Respiratory Medicine* 10, no. 7: 715–724. [https://doi.org/10.1016/S2213-2600\(22\)00169-2](https://doi.org/10.1016/S2213-2600(22)00169-2).
- Niederberger, M., and J. Spranger. 2020. "Delphi Technique in Health Sciences: A Map." *Frontiers in Public Health* 8: 457. <https://doi.org/10.3389/fpubh.2020.00457>.
- Peng, L., J. Li, J. Zhou, et al. 2024. "The Development and Initial Validation of IgG4-related Disease Damage Index: A Consensus Report From Chinese IgG4-RD Consortium." *RMD Open* 10, no. 1: e003938. <https://doi.org/10.1136/rmdopen-2023-003938>.
- Ranse, J., B. Mackie, J. Crilly, et al. 2025. "Strengthening Emergency Department Response to Chemical, Biological, Radiological, and Nuclear Disasters: A Scoping Review." *Australasian Emergency Care* 28: 37–47. <https://doi.org/10.1016/j.auec.2024.09.003>.
- Ruan, Y., S. Song, and Z. Yin, et al. 2022. "Comprehensive Evaluation of Military Training-Induced Fatigue Among Soldiers in China: A Delphi Consensus Study." *Frontiers in Public Health* 10: 1004910. <https://doi.org/10.3389/fpubh.2022.1004910>.
- Salmi, M., D. Atif, D. Oliva, A. Abraham, and S. Ventura. 2024. "Handling Imbalanced Medical Datasets: Review of a Decade of Research." *Artificial Intelligence Review* 57, no. 10: 273. <https://doi.org/10.1007/s10462-024-10884-2>.
- Smith, E. M. D., S. Rasul, C. Ciurtin, et al. 2021. "Limited Sensitivity and Specificity of the ACR/EULAR-2019 Classification Criteria for SLE in JSLE?—Observations From the UK JSLE Cohort Study." *Rheumatology* 60, no. 11: 5271–5281. <https://doi.org/10.1093/rheumatology/keab210>.
- Spranger, J., A. Homberg, M. Sonnberger, and M. Niederberger. 2022. "Reporting Guidelines for Delphi Techniques in Health Sciences: A Methodological Review." *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen* 172: 1–11. <https://doi.org/10.1016/j.zefq.2022.04.025>.
- Uuk, R., A. Brouwer, T. Schreier, N. Dreksler, V. Pulignano, and R. Bommasani. 2024. "Effective Mitigations for Systemic Risks from General-Purpose AI (arXiv:2412.02145)." *arXiv* 1: 1. <https://doi.org/10.48550/arXiv.2412.02145>.
- Vellido, A. 2020. "The Importance of Interpretability and Visualization in Machine Learning for Applications in Medicine and Health Care." *Neural Computing and Applications* 32, no. 24: 18069–18083. <https://doi.org/10.1007/s00521-019-04051-w>.
- Vergni, L., F. Todisco, and B. Di Lena. 2021. "Evaluation of the Similarity Between Drought Indices by Correlation Analysis and Cohen's Kappa Test in a Mediterranean Area." *Natural Hazards* 108, no. 2: 2187–2209. <https://doi.org/10.1007/s11069-021-04775-w>.
- Zon, M., G. Ganesh, M. J. Deen, and Q. Fang. 2023. "Context-Aware Medical Systems Within Healthcare Environments: A Systematic Scoping Review to Identify Subdomains and Significant Medical Contexts." *International Journal of Environmental Research and Public Health* 20, no. 14: 6399. <https://doi.org/10.3390/ijerph20146399>.

## Supporting Information

Additional supporting information can be found online in the Supporting Information section.