# Development, validation, and usage of metrics to evaluate clinical research hypothesis quality

Xia Jing, PhD[1], Yuchun Zhou, PhD[2], James J. Cimino, MD[3], Jay H. Shubrook, DO[4], Vimla L. Patel, PhD[5], Sonsoles De Lacalle, MD, PhD[6], Aneesa Weaver, BS[1], Chang Liu, PhD[7]

[1]College of Behavioral, Social, and Health Sciences, Clemson University, Clemson, South Carolina, USA; [2]Patton College of Education, Ohio University, Athens, Ohio, USA; [3]Informatics Institute, School of Medicine, University of Alabama, Birmingham, Alabama, USA; [4]College of Osteopathic Medicine, Touro University, Vallejo, California, USA; [5]The New York Academy of Medicine, New York, New York, USA; [6]College of Art and Science, California State University Channel Islands, Camarillo, California, USA; [7]Russ College of Engineering and Technology, Ohio University, Athens, Ohio, USA

**Corresponding author:**

Xia Jing, Department of Public Health Sciences, College of Behavioral, Social, and Health Sciences, Clemson University, 519 Edwards Hall, Clemson, SC 29634; xjing@clemson.edu

**Keywords:**

Clinical hypothesis evaluation; metrics; instrument development; validation; clinical research; scientific hypotheses evaluation

**Word count:** 1998 (exclude title, abstract, keywords, tables, table titles, figure titles, and references)

**Abstract (145 words)**

***Objectives****: To develop, validate, and use evaluation instruments to assess the quality of clinical hypotheses generated using secondary data analytic tools.*

***Materials and Methods****: The development of the evaluation metrics went through iterative stages, including literature review, internal and external validations, testing, and continuous revisions through feedback from seven experts. We conducted two experiments to determine a brief version of the instrument for a gateway evaluation.*

***Results****: The brief version of the instrument contained validity, significance, and feasibility based on experiment results. The comprehensive metrics also included novelty, clinical relevance, potential benefits and risks, ethicality, testability, clarity, and interestingness based on literature review, internal and external validations.*

***Conclusion****: The developed and validated brief and comprehensive versions of the clinical hypotheses evaluation metric can provide standardized and generic measurements for clinical research hypotheses and allow clinical researchers to prioritize their research ideas systematically, objectively, and consistently.*

## INTRODUCTION

A hypothesis is an educated guess or statement about the relationship between two or more variables [1,2]. The hypothesis-generation process is critical and decisive in determining the significance of a clinical research or scientific project. Although there is much progress in scientific thinking, reasoning, and analogy [3-8], all of which are critical skills in hypothesis generation, there is limited knowledge regarding the scientific hypothesis generation process itself, including how to facilitate the process, especially in a clinical research context. Many data science researchers believe that secondary data analytic tools can facilitate the hypothesis generation, e.g., [9]. Nevertheless, there is a lack of studies demonstrating how a secondary data analysis tool can facilitate this process in clinical research or the extent of such facilitation. We developed a Visual Interactive Analytic tool for filtering and summarizing large health Data Sets coded with hierarchical terminologies (VIADS [10], https://www.viads.info) to assist clinical researchers with generating hypotheses. Visual examples of VIADS include hierarchical graphs, bar charts, and 3D plots. Users can obtain expanded information via interactive features, change graph layouts, and move and export graphs.

To explore the clinical researchers' hypothesis generation processes, we conducted study sessions in which researchers (i.e., participants) generated hypotheses using the same dataset, within a 2-hour timeframe, with or without VIADS [11]. This was a 2 × 2 study design (with and without VIADS by experienced and inexperienced clinical researchers). Each scientific hypothesis (a total of 19 during the pilot study sessions and 227 during the formal study sessions) generated by the participants in the study [12,13] was assessed by an expert panel using the same metrics. The aggregated assessment results were used to detect the differences in the hypotheses generated by the participants [12]. To have a reliable, consistent, and generic assessment of scientific hypotheses, a reliable and valid tool is required to accurately, consistently, and conveniently evaluate the quality of a clinical research hypothesis [14]. Here we introduce the approach we used to develop and validate the metrics and usage of the metrics.

## METHODS

Development of the metrics went through a series of iterative stages (Figure 1) [15-17]. One author (XJ, a medical informatics researcher) reviewed the literature and drafted the metrics. Then two authors (XJ and YCZ, a research methodologist) discussed the outlined metrics, formulated the initial metrics, and revised the metrics after all concerns were addressed. The adjusted metrics were distributed to the research team for feedback, which constituted the second level of internal validation among the entire team. After completing the internal validation, we conducted an iterative external validation process by engaging four invited clinical research experts (Appendix 1). The internal and external validation processes followed a revised Delphi method [18-22], which included transparent and open discussions (via email) among the research team and the anonymous survey about the evaluation items. The external validation consisted of three steps, initial

validation, experimental evaluations by using metrics to assess hypotheses generated during the 2 × 2 study, and refinement based on the feedback and results of the experimental evaluations (Figure 2). A 10-item evaluation instrument was formulated from the development and validation processes. A second survey (Appendix 2) that concerns the current version of the final metrics will be continuously revised by incorporating the comments from the evaluation results.

**Experimental evaluation 1**

In experimental evaluation 1, we performed validation analysis for the 10 evaluation items (without subitems) using 19 hypotheses generated via pilot studies of the 2 × 2 study. These hypotheses were randomly assigned into two Qualtrics surveys (10 and 9 hypotheses). Four expert panel members and three senior advisors with a medical or methodology background responded to the surveys. The inter-rater agreement of the seven experts' rating on the 19 hypotheses was analyzed using the intra-class correlation (ICC). We used descriptive statistics to analyze the results of the survey. Based on the mean results from experimental evaluation 1, we identified the best and worst examples (Figure 3) of hypotheses, which were used in experimental evaluation 2.

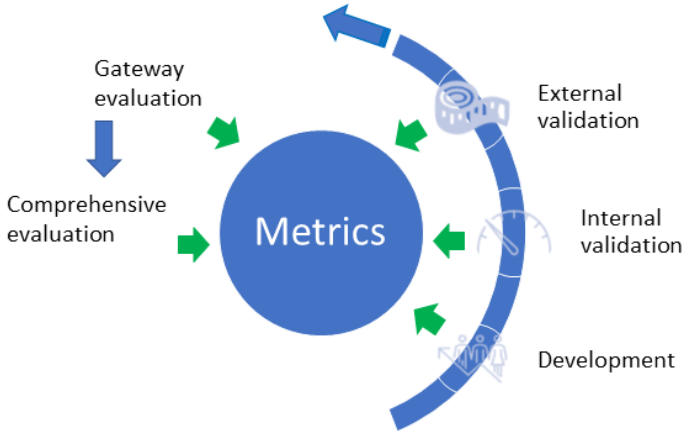


**Figure 1.** Development, validation, and usage of metrics to assess clinical research hypotheses (the blue arrows indicate the development stages of metrics; the green arrows indicate the feedback from each stage to metrics)

**Experimental evaluation 2**

Experimental evaluation 2 included 30 randomly selected hypotheses from the study sessions using the 10-item evaluation instrument. In the instructions, we provided the best and worst examples of hypotheses based on the experimental evaluation 1 results and set a screening item: validity (Figure 3). If a statement is not a hypothesis, then further evaluation is unnecessary. If three or more experts scored at 1 (lowest rating) in validity for any of the hypotheses, it was removed from the following analysis. ICC analysis was performed to examine the consistency of the seven experts' ratings on the valid hypotheses using the ten items. A valid hypothesis

means two or fewer experts assessed 1 (the lowest rating) in its validity. A paired *t*-test analysis was used to compare the evaluation results using the 10-item or a brief version of the instrument.

All steps mentioned above (initial draft development, internal validation, external validation, refinement, and revisions in between the steps) were conducted iteratively using quantitative and qualitative approaches (e.g., Qualtrics surveys, emails, additional phone calls, and virtual conferences). Iterative evaluations of the instrument validation process (i.e., before experts used the instrument to conduct the experimental evaluation 1) included a 5-point Likert scale and three additional options of unable to assess, unnecessary subitem, or use this item only. The evaluation instrument used in experimental evaluations 1 and 2 included a 5-point Likert scale and an option of not applicable. The predetermined inclusion criteria of the expert panel members can be found in our prior publication [11]. This study was approved by the Ohio University Institutional Review Board (18-X-192) and Clemson University Institutional Review Board (IRB2020-056).

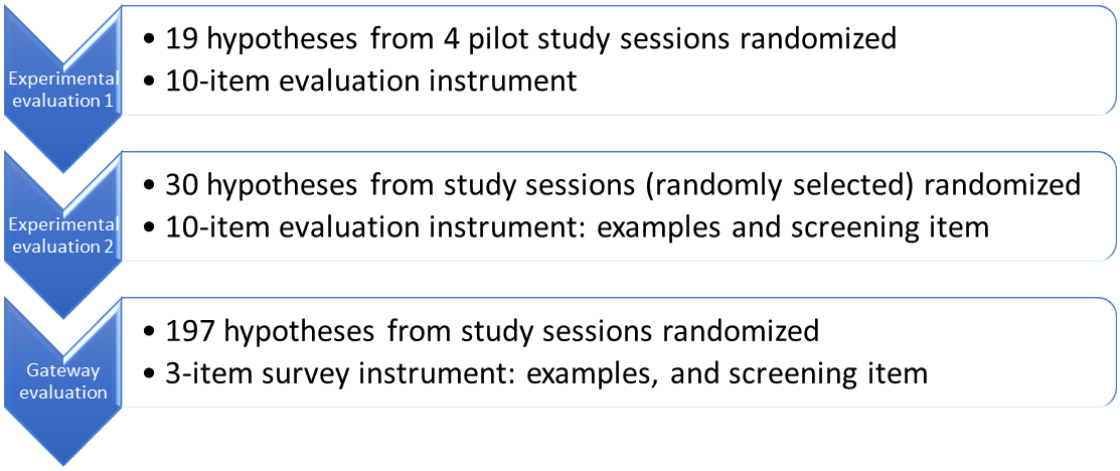


**Figure 2**. Refinement process of the clinical research hypotheses evaluation instrument

## RESULTS

We present a comprehensive (10-item and 39-subitem, Appendix 3) and a brief versions (3-item, Figure 3) of instrument to assess clinical researc hypotheses and the evidence generated from expereimental evaluations. Most measurements for evaluating clinical research hypotheses from literature [1,2,9,23-33] include the following ten dimensions: ***validity***, ***significance***, ***novelty***, ***clinical relevance***, ***potential benefits and risks***, ***ethicality***, ***feasibility***, ***testability***, ***clarity***, and ***researcher interest level***. We developed 39 sub-items to measure each dimension comprehensively and unambiguously (Table 1). The quality of each item was measured using a 5-point Likert scale. Table 1 lists all the evaluation items and subitems and how they were used to evaluate the clinical research hypotheses. Table 2 presents two example hypotheses and their evaluation results among all evaluators when using 3-item instrument (Figure 3). Appendix 4 presents the brief version of the 10-item evaluation instrument without the subitems, which was used in the experimental evaluation 2.

**Table 1.** Evaluation items and subitems in the metrics used to assess the scientific hypotheses in clinical research

| **10 Evaluation items** | **39 Subitems (39)** | **Note** |
|---|---|---|
| Clarity | | |
| | Clear purposes | The hypothesis is clear in each aspect (i.e., subitems) evaluated on a 5-point Likert scale. |
| | Clear, focused groups | |
| | Specified variables | |
| | Specified relationships among variables | |
| | Overall clear | |
| Clinical relevance | | |
| | Impact on current clinical practice | To test if the hypothesis has the potential to have a significant impact on each of these aspects (i.e., subitems), evaluated on a 5-point Likert scale. |
| | Impact medical knowledge | |
| | Impact health policy | |
| | Overall clinically relevant | |
| Ethicality | | |
| | No ethical concerns | When conducting a study to test a given hypothesis, there are no ethical concerns (regarding stakeholders and conduction). Consider using *binary options* instead of a 5-point Likert scale. |
| | Trade my place with a participant if eligible | |
| | Overall, an ethical study to test | |
| Feasibility | | |
| | Regarding needed costs | To test if the hypothesis is feasible regarding the available resources and scope of the work, evaluated on a 5-point Likert scale. |
| | Regarding needed time | |
| | Regarding the scope of the work | |
| | Overall feasible | |
| Interestingness | | |
| | It interests me | The researcher should be able to find interested collaborators easily in the field; consider using *binary options* instead of a 5-point Likert scale. |
| | I will pursue it if possible. | |
| | Overall an interesting idea | |
| Novelty | | |
| | Leads to innovation in medical practice | To test if the hypothesis has the potential to lead to innovations in each of these aspects (i.e., subitems), evaluated on a 5-point Likert scale. |
| | This leads to new methodologies for clinical research. | |
| | It may alter previous findings. | |

| | | |
|---|---|---|
| | Leads to novel medical knowledge | |
| | This leads to new findings, which can be incremental. | |
| | Overall novel | |
| Potential benefits and risks | | |
| | Significant benefits | To test if the hypothesis has the potential to provide significant benefits over risks to stakeholders, consider using *binary options* instead of a 5-point Likert scale. |
| | No or tolerable risks | |
| | The overall benefits outweigh the risks. | |
| Significance | | |
| | Addressing established medical needs | To test if the hypothesis has the potential to have an impact on each of these aspects (i.e., subitems), evaluated on a 5-point Likert scale. |
| | Impact future direction of the field | |
| | Impact on the target population | |
| | Impact the cost and benefit | |
| | Overall significant | |
| Testability | | |
| | It can be tested in an ideal setting. | The hypothesis can be tested, regardless of feasibility, and evaluated on a 5-point Likert scale. |
| | Adequate number of patients to choose from | |
| | Overall testable | |
| Validity | | |
| | Scientific validity | The hypothesis is scientifically and clinically valid, evaluated on a 5-point Likert scale. |
| | Clinical validity | |
| | Overall valid | |

Note: Validity, significance, and feasibility, denoted in a green background, were used in a breif version of instrument to conduct gateway evaluations for the hypotheses generated.

In experimental evaluation 1, the experts' evaluation scores for the 19 hypotheses across the ten criteria were averaged, and none of the ten criteria could achieve a moderate ICC coefficient (>0.50). Therefore, we conducted experimental evaluation 2 and set validity as a screening item and provided one best and one not so good example hypotheses in the instructions of the experimental evaluation 2.

In the experimental evaluation 2 result analysis, we checked the results of the screening item first. The valid sample size included 17 hypotheses in experimental evaluation 2. We then checked the inter-rater agreement of the 17 hypotheses using ICC analyses. Half of the ten criteria achieved a moderate ICC value (.50–.75). Based on the ICC results and qualitative evaluation of the ten criteria, we decided to retain three criteria (i.e., validity, significance, and feasibility) for a shortened version of the evaluation instrument.

**Table 2** Example hypotheses and evaluation results by using the 3-item instrument

| Hypothesis | 3 Evaluation items | R1 | R2 | R3 | R4 | R5 | R6 | R7 | Item means | SD | Hypothesis mean ± 95% confidence interval |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **H1** | **Validity** | 5 | 5 | 5 | 3 | 4 | 3 | 4 | 4.14 | 0.83 | 4.0 ± 0.35 |
| | **Significance** | 4 | 4 | 4 | 3 | 3 | 2 | 4 | 3.43 | 0.73 | |
| | **Feasibility** | 5 | 4 | 4 | 4 | 5 | 5 | 4 | 4.43 | 0.49 | |
| **H2** | **Validity** | NA | 3 | 4 | 3 | 3 | 3 | 1 | 2.83 | 0.90 | 2.64 ± 0.52 |
| | **Significance** | NA | 3 | 4 | 4 | 2 | 3 | 2 | 3 | 0.96 | |
| | **Feasibility** | NA | 3 | 2 | 3 | 2 | 1 | 1 | 2 | 0.82 | |

Note: hypothesis 1: Patients who have hypertension between 2005 and 2015, do hypertension patients have a higher rate of morbid obesity (ICD9 codes: 27801) in 2015 than in 2005?; hypothesis 2: Whether the changes in packed food consumption caused the increase in diabetes (ICD9 code: #250) from 2005 (case counts: 774 ) to 2015 (case counts:1281) at the zip code level. R1: reviewer 1; NA: not applicable, i.e., an evaluator is unable to aseess the item or a hypothesis is invalid and all following items are not assessed.

The paired *t*-test indicated that no significant difference (t = 1.74, p = .13) between the ratings using the 3-item instrument and the 10-item instrument. Figure 3 shows the 3-item evaluation instrument used for gateway evaluations, including best and worst examples. Appendix 5 provides a summary of the methods, steps and the correspoidning results of each step.

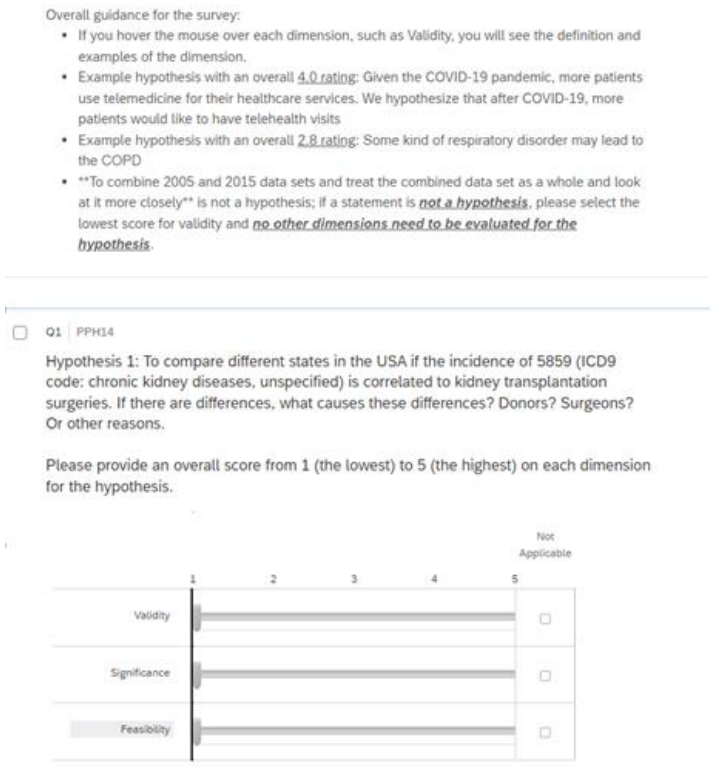


**Figure 3**. Three-item evaluation instrument for initial clinical research hypothesis screening and evaluation

## DISCUSSION

Hypothesis generation is a highly sophisticated cognitive process. Not all information use and processes are conscious or explicit. Our study explored the process of scientific hypothesis generation using the same clinical data sets to determine whether a secondary data analytic tool could facilitate the process. Establishing the evaluation metrics was the first step and was the critical foundation for the overall study and understanding of the entire process. The comprehensive and objective aspects were given more weight when developing the metrics. In our studies, we noticed that clinical researchers generated a few to over a dozen hypotheses [12,13]. But, not all hypotheses were of high quality. Therefore, it was not conducive to use the experts' time to comprehensively evaluate every hypothesis that was generated during the study sessions.

Furthermore, using the entire set of metrics to evaluate each generated hypothesis may not be necessary. Thus, we used "gateway" evaluations as a filter to identify the higher-quality hypotheses. During the comprehensive evaluation, the experts will then be able to evaluate the higher-quality hypotheses more carefully, thoroughly, and comprehensively. Therefore, we used validity as a screening item and added the "not a hypothesis" option in the initial assessment, enlightened by the experimental evaluation 1 results.

The results of experimental evaluation 2 aided in developing a brief evaluation instrument with the 3 items, which are used to evaluate the rest of the hypotheses generated by the participants during the gateway evaluation (Figure 1). From the ICC analysis in experimental evaluation 2, feasibility, testability, and clarity have the highest ICC values among the ten items. However, empirically, we highly prioritize validity, significance, and novelty. Combining our experience and the statistical testing results, we developed two options: validity, significance, and feasibility; validity, significance, clinical relevance, and feasibility. The testing results indicated that both were valid options. Thus, we determined the 3-item evaluation instrument for operational purposes. We used our experience and the statistical testing results to make the decision.

Meanwhile, we noticed that there are negative ICC values in ethicality, potential benefits and risks, and interestingness. The results indicated that it might be hard to reach a consensus on these items. We recommend these three items change to a binary (yes/no) category instead of a 5-point Likert scale to simplify the evaluation and improve the agreement among the evaluators.

During the external validation, one major result was to add "not applicable" as an option to the evaluation instrument under each item and subitem. Considering the different backgrounds of expert panel members, this additional option helped them to simplify the evaluation process. Comparing the statistical results, we noticed a significant improvement in experimental evaluation 2, mainly due to the examples of the best and worst hypotheses, which might assist evaluators in calibrating their expectations. Furthermore, we reminded the evaluators that some statements were not hypotheses, i.e., we used validity as the screening items. The results

of experimental evaluation 2 are based on 17 valid hypotheses. The 13 invalid hypotheses have three or more expert panel members who evaluated them as a 1 (the lowest score) in validity.

Although the evaluation of a particular hypothesis by an expert can be subjective, we used examples of the best and worst hypotheses to assist experts in calibrating their expectations more accurately. The inclusion of seven expert members balances the subjectivity and provides a more consistent evaluation using the same instrument. In addition, we used objective measures, e.g., the number of hypotheses generated and the average time spent on each hypothesis, and randomized the hypotheses during the assessment. These strategies helped the expert panel to provide more consistent evaluations and allowed us to accurately conclude the quality of the hypotheses.

## CONCLUSION

The metrics and instruments developed in this study can benefit clinical researchers in evaluating their hypotheses more comprehensively, consistently, and efficiently before launching a research project, as well as providing valid instruments for the peer-review process in clinical research. Our results provide an evidence-based brief version (validity, significance, and feasibility) and a comprehensive version of the evaluation items (validity, significance, feasibility, novelty, clinical relevance, testability, clarity, ethicality, potential benefits and risks, and interesting to others) to assess clinical research hypotheses. Metrics can be used to standardize the process and provide a consistent instrument for this highly sophisticated cognitive process.

**Appendix 1**: Initial survey instrument used for external validation of the evaluation items

**Appendix 2**: Medium survey instrument used for external validation of the evaluation items

**Appendix 3**: Evaluation instrument of 10-item with subitems (full-version) to evaluate the scientific hypotheses in clinical research

**Appendix 4:** Evaluation instrument of 10-item (without subitems) to evaluate the scientific hypotheses in clinical research

**Appendix 5**: Summary of methods, steps, and corresponding results of development and validation of metrics in assessing clinical hypotheses

## Acknowledgments and funding

The project is supported by a grant (R15LM012941) from the National Library of Medicine and partially under a grant (P20GM121342) from the National Institute of General Medical Sciences of the United States National Institutes of Health.

## Competing interests

Non to disclose.

## References

1. Supino P, Borer J. Principles of research methodology: A guide for clinical investigators. 2012
2. Parahoo A. Nursing research: Principles, Process & issues. 1997
3. *The Oxford handbook of thinking and reasoning*. New York, NY, US: Oxford University Press, 2012.
4. Patel VL, Arocha JF, Zhang J. Chapter 30: Thinking and Reasoning in Medicine. In: Holyoak KJ, Morrison RG, eds. The Cambridge Handbook of Thinking and Reasoning. New York: Cambridge University Press, 2005:727-50.
5. Joseph G-M, Patel VL. Domain knowledge and hypothesis generation in diagnostic reasoning. *Medical Decision Making* 1990;**10**:31-46.
6. Kaufman DR, Patel VL, Magder SA. The explanatory role of spontaneously generated analogies in reasoning about physiological concepts. *International Journal of Science Education* 1996;**18**(3):369-86. doi: 10.1080/0950069960180309
7. Arocha J, Patel V, Patel Y. Hypothesis generation and the coordiantion of theory and evidence in novice diagnostic reasoning. *Medical Decision Making* 1993;**13**:198-211.
8. Patel V, Groen G. Knowledge Based Solution Strategies in Medical Reasoning. *Cognitive Sci* 1986;**10**:91-116. doi: 10.1207/s15516709cog1001_4
9. Spangler S. Accelerating discovery : mining unstructured information for hypothesis generation. 2016
10. Jing X, Emerson M, Masters D, et al. A visual interactive analysis tool for filtering and summarizing large data sets coded with hierarchical terminologies (VIADS). *BMC Med Inform Decis Mak* 2019;**19**(31) doi: https://doi.org/10.1186/s12911-019-0750-y
11. Jing X, Patel VL, Cimino JJ, et al. The Roles of a Secondary Data Analytics Tool and Experience in Scientific Hypothesis Generation in Clinical Research: Protocol for a Mixed Methods Study. *JMIR Res Protoc* 2022;**11**(7):e39414. doi: 10.2196/39414
12. Jing X, Patel VL, Cimino JJ, et al. A visual analytic tool to assist hypothesis generation process in clinical research—A utility and usability study of VIADS using mixed-methods. *JMIR (Under review)* 2023 doi: https://preprints.jmir.org/preprint/44644
13. Draghi B, Ernst M, Patel V, et al. Number of scientific hypotheses and time needed in a 2-hour study session among inexperienced clinical researchers—preliminary results. AMIA Summit 2023; 2023 Mar 13-16, 2023; Seattle, Washington.
14. Jing X, Zhou YC, Cimino JJ, et al. Development and preliminary validation of metrics to evaluate data-driven clinical research hypotheses. AMIA 2022; 2022 Nov 5-9, 2022; Washington DC.
15. Oates BJ. *Researching information systems and computing*. London: SAGE publications, 2006.
16. Wisdom J, Creswell JW. Mixed Methods: Integrating Quantitative and Qualitative Data Collection and Analysis While Studying Patient-Centered Medical Home Models. Secondary Mixed Methods: Integrating Quantitative and Qualitative Data Collection and Analysis While Studying Patient-Centered Medical Home Models 2013. https://pcmh.ahrq.gov/page/mixed-methods-integrating-quantitative-and-qualitative-data-collection-and-analysis-while.
17. Kelley-Quon, I L. Surveys: Merging Qualitative and Quantitative Research Methods. *Seminars in pediatric surgery* 2018;**27**(6):361–66.
18. RAND Corporation. Delphi Methods. Secondary Delphi Methods. https://www.rand.org/topics/delphi-method.html.
19. Chitu Okoli, Suzanne D. Pawlowski. The Delphi method as a research tool: an example, design considerations and applications. *Information & Management* 2004;**42**(1):15-29. doi: https://doi.org/10.1016/j.im.2003.11.002

20. Hohman N, Cote M, Brand J. Research Pearls: Expert Consensus Based Evidence Using the Delphi Method. *Arthroscopy* 2018;**34**:3278-82. doi: 10.1016/j.arthro.2018.10.004
21. Khodyakov D, Grant S, Denger B, et al. Using an Online, Modified Delphi Approach to Engage Patients and Caregivers in Determining the Patient-Centeredness of Duchenne Muscular Dystrophy Care Considerations. *Med Decis Making* 2019;**39**(8):1019-31. doi: 10.1177/0272989x19883631
22. Hill CE, Knox S. *Essentials of Consensual Qualitative Research*: American Psychological Association, 2021.
23. Hicks CM. Research methods for clinical therapists: Applied project design and analysis. 1999
24. Glasser SP. Essentials of clinical research. 2014
25. Farrugia P, Petrisor B, Farrokhyar F, Bhandari M. Research questions, hypotheses and objectives. *J Can Chir* 2010;**50**
26. Biesecker L. Hypothesis-generating research and predictive medicine. *Genome Res* 2013;**23**:1051-53.
27. Friedman LM, Furberg CD, DeMets DL, Reboussin DM, Granger CB. *Fundamentals of Clinical Trials*: Springer International Publishing, 2015.
28. Portney LG. *Foundations of Clinical Research: Applications to Evidence-based Practice*: F.A. Davis, 2020.
29. Hulley S, Cummings S, Browner W, Grady D, Newman T. Designing clinical research. 2013
30. Richesson RL, Andrews JE. Clinical Research Informatics. 2012
31. Embi PJ, Payne PRO. Clinical research informatics: challenges, opportunities and definition for an emerging domain. *JAMIA* 2009;**16**:316-27.
32. Gallin JI, Ognibene FP, Ognibene FP. *Principles and Practice of Clinical Research*. Burlington, UNITED STATES: Elsevier Science & Technology, 2007.
33. Pruzan P. *Research Methodology: The Aims, Practices and Ethics of Science*: Springer International Publishing Switzerland, 2016.

**Figure legends**

**Figure 1.** Development, validation, and usage of metrics to assess clinical research hypothesis

**Figure 2.** Refinement process of the clinical research hypotheses evaluation instrument

**Figure 3.** Three-item evaluation instrument for initial clinical research hypothesis screening and evaluation