



STAGER checklist: Standardized testing and assessment guidelines for evaluating generative artificial intelligence reliability

Jinghong Chen^{1*}, Lingxuan Zhu^{1*}, Weiming Mou^{1,2*}, Anqi Lin^{1*}, Dongqiang Zeng³, Chang Qi⁴, Zaoqu Liu^{5,6}, Aimin Jiang⁷, Bufu Tang⁸, Wenjie Shi⁹, Ulf D Kahlert⁹, Jianguo Zhou^{10,11,12}, Shipeng Guo¹³, Xiaofan Lu¹⁴, Xu Sun¹⁵, Trunghieu Ngo¹⁵, Zhongji Pu¹⁶, Baolei Jia¹⁶, Che Ok Jeon¹⁷, Yongbin He^{18,19}, Haiyang Wu^{20,21}, Shuqin Gu²², Wisit Cheungpasitporn²³, Haojie Huang^{24,25,26}, Weipu Mao^{27,28}, Shixiang Wang²⁹, Xin Chen³⁰, Loïc Cabannes¹⁵, Gerald Sng Gui Ren^{31,32}, Iain S Whitaker^{33,34}, Stephen Ali^{33,34}, Quan Cheng^{35,36#}, Kai Miao^{37,38#}, Shuofeng Yuan^{39,40#}, Peng Luo^{1#}

¹Zhujiang Hospital, Southern Medical University, China

²Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, China

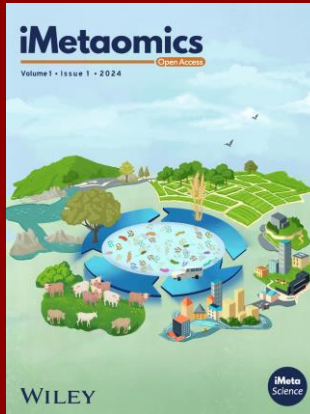
³⁵Xiangya Hospital, Central South University, Changsha 410008, China

³⁷Faculty of Health Sciences, University of Macau, Macau SAR 999078, China

³⁸MoE Frontiers Science Center for Precision Oncology, University of Macau, Macau SAR 999078, China

³⁹The University of Hong Kong-Shenzhen Hospital, China

⁴⁰Li Ka Shing Faculty of Medicine, The University of Hong Kong, China



Jinghong Chen, Lingxuan Zhu, Weiming Mou, Anqi Lin, et al. 2024. STAGER checklist: Standardized testing and assessment guidelines for evaluating generative artificial intelligence reliability. *iMetaOmics* 1: e1.

<https://doi.org/10.1002/imo2.1>



Introduction

Generative artificial intelligence (AI) holds immense potential in medical applications. Numerous studies have explored the efficacy of various generative AI models within healthcare contexts, but there is a lack of a comprehensive and systematic evaluation framework. Given that some studies evaluating the ability of generative AI for medical applications have deficiencies in their methodological design, standardized guidelines for their evaluation are also currently lacking. In response, our objective is to devise standardized assessment guidelines tailored for evaluating the performance of generative AI systems in medical contexts.



Introduction

generative AI capabilities in medicine

Web of Sciences
Cochrane Library
PubMed
Google Scholar

01

02

multidisciplinary team

comprising experts in life sciences, clinical medicine, medical engineering, and generative AI users

03

STAGER checklist

A list of 32 projects covering key assessment aspects of generative AI in healthcare applications

The checklist is designed to encompass the critical evaluation aspects of generative AI in medical applications comprehensively. This checklist, and the broader assessment framework it anchors, address several key dimensions, including question collection, querying methodologies, and assessment techniques. Our framework furnishes a standardized and systematic approach for research involving the testing of generative AI's applicability in medicine. It enhances the quality of research reporting and aids in the evolution of generative AI in medicine and life sciences.



Publication records of a PubMed search using 'ChatGPT' as the keyword

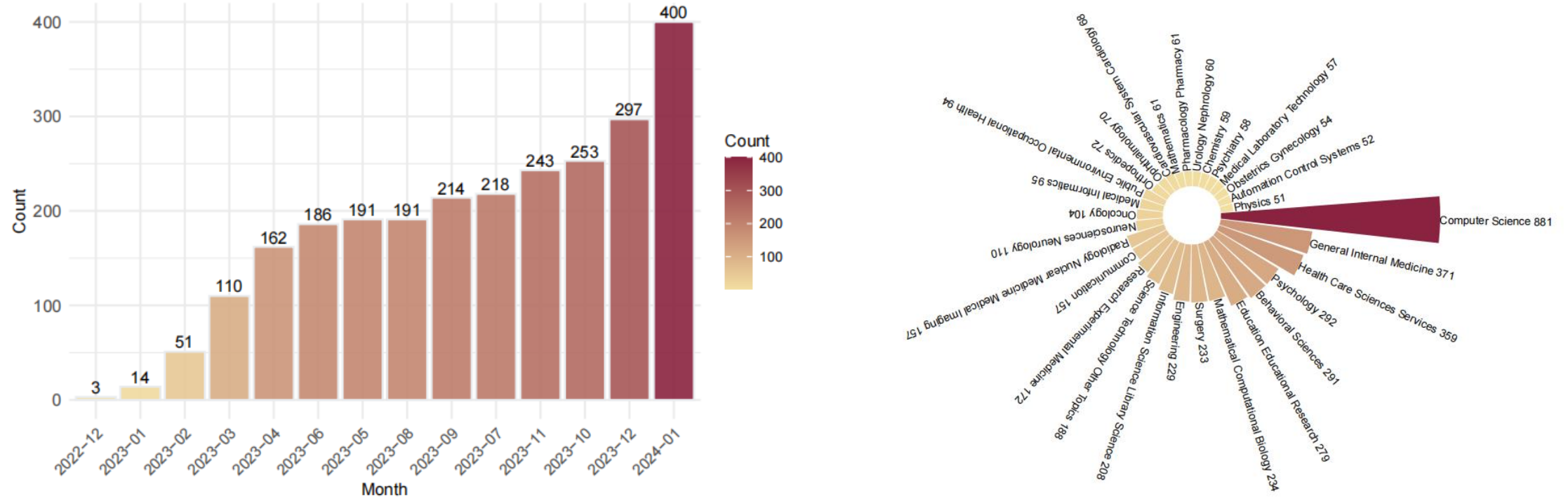
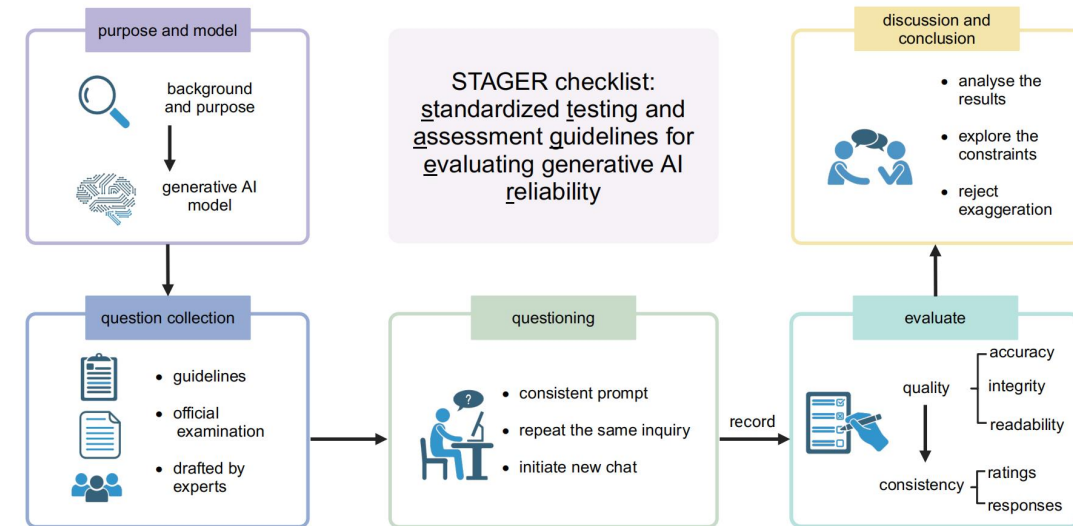


Figure 1 Publication records of a PubMed search using 'ChatGPT' as the keyword.



Highlights

- This work formulates the Standardized Testing and Assessment Guidelines for Evaluating Generative Artificial Intelligence (AI) Reliability (STAGER) checklist, a 32-item framework offering standardized assessment guidelines tailored for evaluating generative AI systems in medical and life science contexts.
- It is consisting of key aspects including question collection, querying approaches, and assessment techniques.
- It enhances research quality and facilitates advances in this emerging field.





Workflows of the STAGER checklist

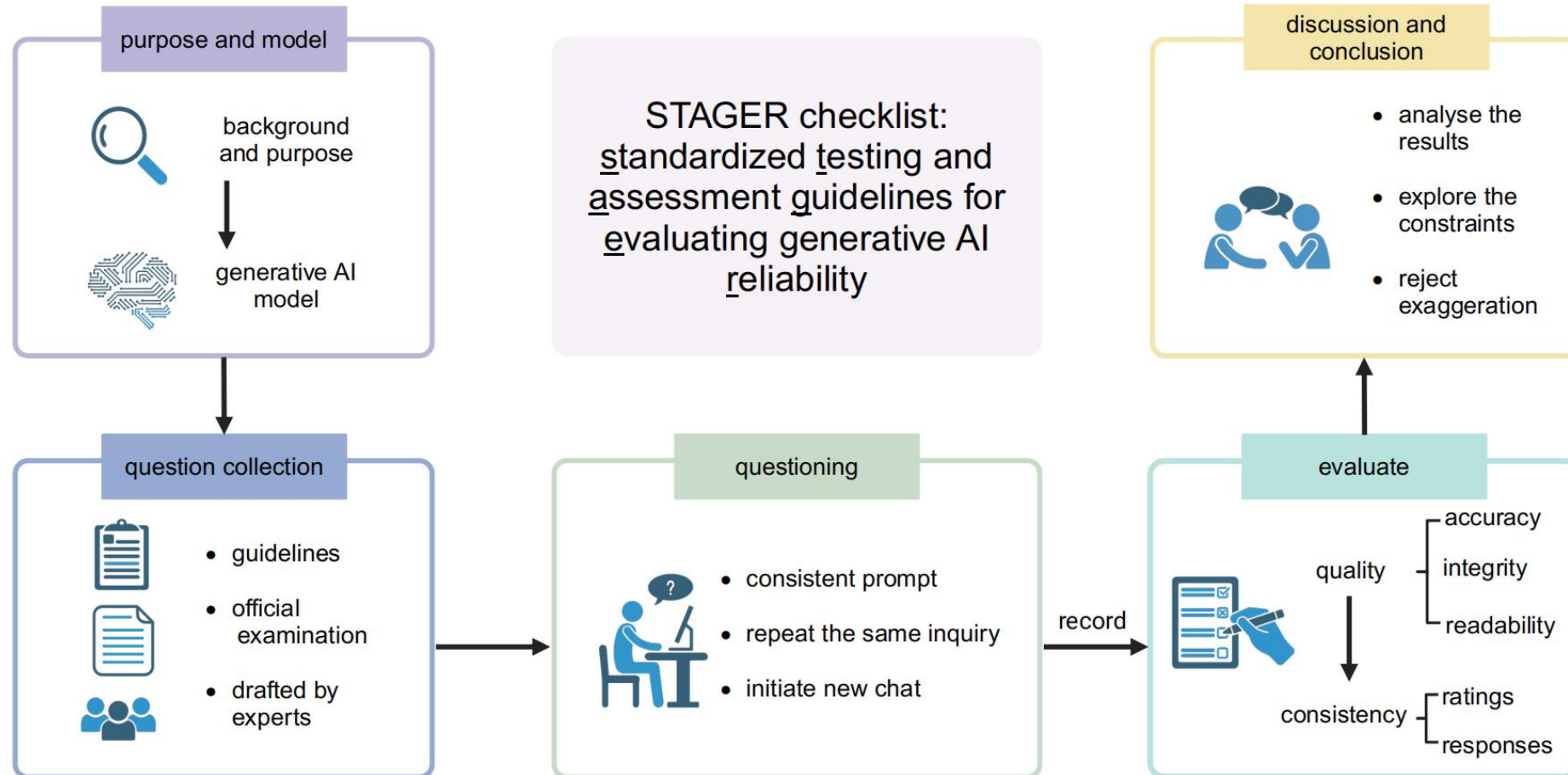


Figure 2 Schematic diagram outlining key components of the STAGER checklist for evaluating generative AI systems.



Details of the STAGER checklist

For a concise checklist without detailed explanations, please refer to Table S or visit GenAIMed.org.

Evaluations and explanations of generative AI for medical applications.			
Section/Topic	Item No	Recommendation	Explanations
Title	1	Identify the report as an article related to the research that evaluates generative AI's applicability in medicine.	Provide the reader with an initial understanding of the nature of the text
	2	State the purpose of the research, the generative AI model used and its version, the source of the questions, methods, results, and conclusions.	Lay the foundation for readers to quickly understand the study and facilitate other researchers to critically analyze the design and results of this research.
Introduction			
Justification	3	Review existing relevant information and explain the background of the study.	Enable readers to grasp the central theme of the article.
Objectives	4	State specific objectives, including the generative AI model used and its version, the training set used for generative AI, the source of the questions, the nature of research, and the limitations.	Provide the necessary framework for readers to understand the article.
Methods			
Question Collection	5	Select the professional questions from guidelines, official examination question banks, and high-frequency issues found via search engines like Google, or drafted by experts, ensuring that the questions cover specific subfields of medicine.	For guidelines or question banks, questions can be either manually selected or extracted using software, while using an API to select questions can reduce subjective errors and make more sense for the entire dataset. When selecting questions from search engines, researchers may opt for frequently occurring ones. If the questions are drafted by experts, the experts need to have authority and experience in the relevant field.
	6	Ensure the questions are representative in terms of difficulty, type, and professionalism.	Enhance the universality of the study.
	7	Describe how the questions were collected, the number of questions, whether the questions were pre-screened, the conditions of the screening, the modality of the input as well as the relevant format.	Input modes such as text, image, sound, video input, etc., and related attributes (e.g., image resolution).
Agent	8	Record the model used, the version of the generative AI, and customized parameters such as temperature parameter, if applicable. State the strengths and weaknesses of the current version used and the rationale for assessing it.	The model used and the version of the generative AI may have a significant influence on the result. Temperature is a parameter influencing text generation randomness. Higher temperatures yield more diverse and novel outputs, with increased unpredictability and potential inaccuracies. Lower temperatures produce consistent, predictable text closely aligned with training data but might lack creativity.
	9	If intend to report them as a functional series, it is recommended to report the relationship between model versions (eg. whether it is a simple upgrade; and if not, it is recommended to report the horizontal comparison results)	Reporting the relationship between model versions clarifies the evolution of the technology, helping users understand improvements or changes. Additionally, providing horizontal comparison results aids in comprehending the distinct capabilities and applications of each version.
Questioning	10	Use a consistent prompt with identically formatted patterns and provide the full prompt in the article.	Reduce the objective differences introduced by different questioning methods and the impact of such differences on the quality of answers. Providing the full prompt in the article ensures that the study is transparent and reproducible.
	11	Ask the same question multiple times and record each response.	Generative AI, known for delivering varied responses to identical queries, necessitates repeated questioning to gauge its consistency.
	12	Indicate whether the question is open-ended or multiple-choice.	Subjective and objective questions are assessed differently.
	13	Initiate a new chat for each question.	Prevent generative AI from being affected by context.
	14	Record the data the responses were collected.	Reduce the impact of performance differences between AI versions and the timing of knowledge updates.
Accuracy	15	Describe any methods employed for scoring accuracy when dealing with subjective questions.	Accuracy refers to the degree to which the response reflects or corresponds to reality or truth.
	16	Compare with reference answers, record the number of correct responses to each question, and calculate the rate of correct answers if you asked objective questions.	The more times the generative AI model responds correctly, the more robust it is considered to be.
Integrity	17	Describe any methods used to assess the integrity between responses.	Integrity refers to whether the response is comprehensive, detailed, and covers relevant information.
Readability	18	Describe any methods used to assess the readability of responses.	Reflect the ease with which a text can be read and understood (e.g. clarity of language, the organization of structure, and grammatical and spelling accuracy).
Reviewers	19	Clarify the composition of reviewers and the rationale for this composition, which is recommended to be more than two experts from varied fields like medicine, artificial intelligence, and interdisciplinary areas, along with stakeholders from ethics, sociology, and user groups.	Ensure the fairness and effectiveness of the evaluation process.
	20	Pay special attention to assessing the implementability of responses.	
	21	Evaluate the consistency across responses to the same question to assess whether the generative AI can steadily provide consistent responses.	
	22	Assess the consistency and reliability of reviewer ratings, avoiding significant differences in the subjective scores among reviewers.	
Results			
Results Selection	23	Describe the results of the search process, from the number of questions collected to the final results, ideally using a flow diagram.	Uncovering its performance in specific subdomains is critical to a deeper understanding of the value and limitations of AI applications in medicine.
Study Characteristics of	24	State all studies included in the analysis and detail their characteristics.	
Results of Individual Studies	25	Present results for accuracy, completeness, and readability for each study, recommending the use of tables or charts for presentation.	
Results of Syntheses	26	Present results of all statistical syntheses conducted, and results of analyses conducted to explore possible causes of heterogeneity among study results.	
Discussion			
Interpretation	27	Analyze the results according to the study objectives.	Comprehensively analyze the performance of generative AI in terms of accuracy, completeness, and readability.
Strengths and Limitations	28	Describe the advantages of the research.	To make the reader understand the importance of the study.
	29	Explore constraints of the research, acknowledging possible origins of partiality or inaccuracy.	Enhance the understanding of the scope, accuracy, and applicability of the research findings
	30	Engage in rational discussion and reject exaggeration.	An honest and rational expression is necessary to maintain academic norms and advance knowledge.
Conclusion	31	Provide a condensed conclusion that summarizes the study's main findings, reiterates its importance, and indicates directions or recommendations for future research.	Provide direction for future research and help promote the further development and application of generative AI in the medical field.
Other Information			
Funding and Sponsorship	32	Provide the origin of financial support and the function of the sponsors for the current investigation, as well as for the initial research if relevant to the foundation of this article.	Maintain the objectivity and transparency of the research.

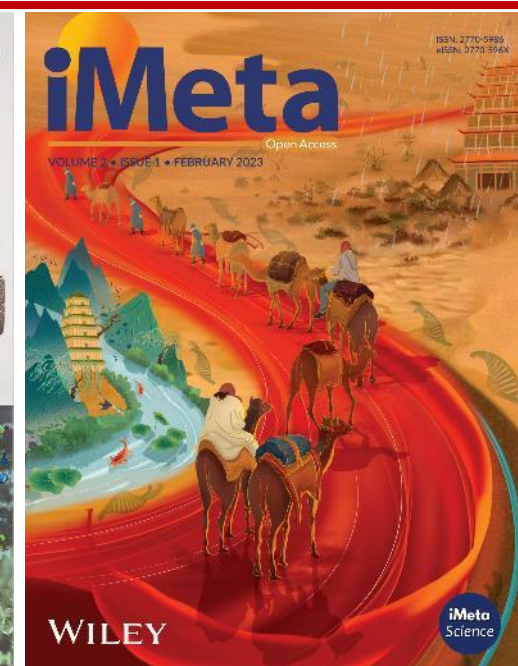


Summary

- ❑ The assessment framework delineated in these guidelines introduces a standardized and systematic method for evaluating generative AI research in medical applications, with an emphasis on elevating the quality of research reports.
- ❑ By providing a clear set of criteria for evaluation, it addresses the need for transparency and rigor in AI research, which is crucial in a field where accuracy and dependability are paramount.
- ❑ This framework is expected to foster academic collaboration and intellectual exchange, creating a fertile ground for cross-disciplinary partnerships, ensuring that it remains cutting-edge, relevant, and aligned with the ever-changing landscape of medical science.
- ❑ STAGER listings and more details are available at GenAIMed.org

Jinghong Chen, Lingxuan Zhu, Weiming Mou, Anqi Lin, et al. 2024. STAGER checklist: Standardized testing and assessment guidelines for evaluating generative artificial intelligence reliability. *iMetaOmics* 1: e1.

<https://doi.org/10.1002/imo2.1>



“***iMeta***” is an open-access Wiley partner journal launched by iMeta Science Society consist of scientists in bioinformatics and metagenomics world-wide. iMeta aims to promote microbiome, and bioinformatics research by publishing research, methods/protocols, and reviews. The goal is to publish high-quality papers (top 10%, IF>20) targeting a broad audience. Unique features include video submission, reproducible analysis, figure polishing, bilingual, and promotion by social media with 500,000 followers. Since 2022 have been published 160 papers and cited > 2300 times. Index by [ESCI](#), [Google Scholar](#), [DOAJ](#) and [Scopus](#).



Society: <http://www.imeta.science>

Publisher: <https://wileyonlinelibrary.com/journal/imeta>

Submission: <https://wiley.atyponrex.com/journal/IMT2>



office@imeta.science



[Promotion Video](#)



[iMetaScience](#)



[iMetaScience](#)