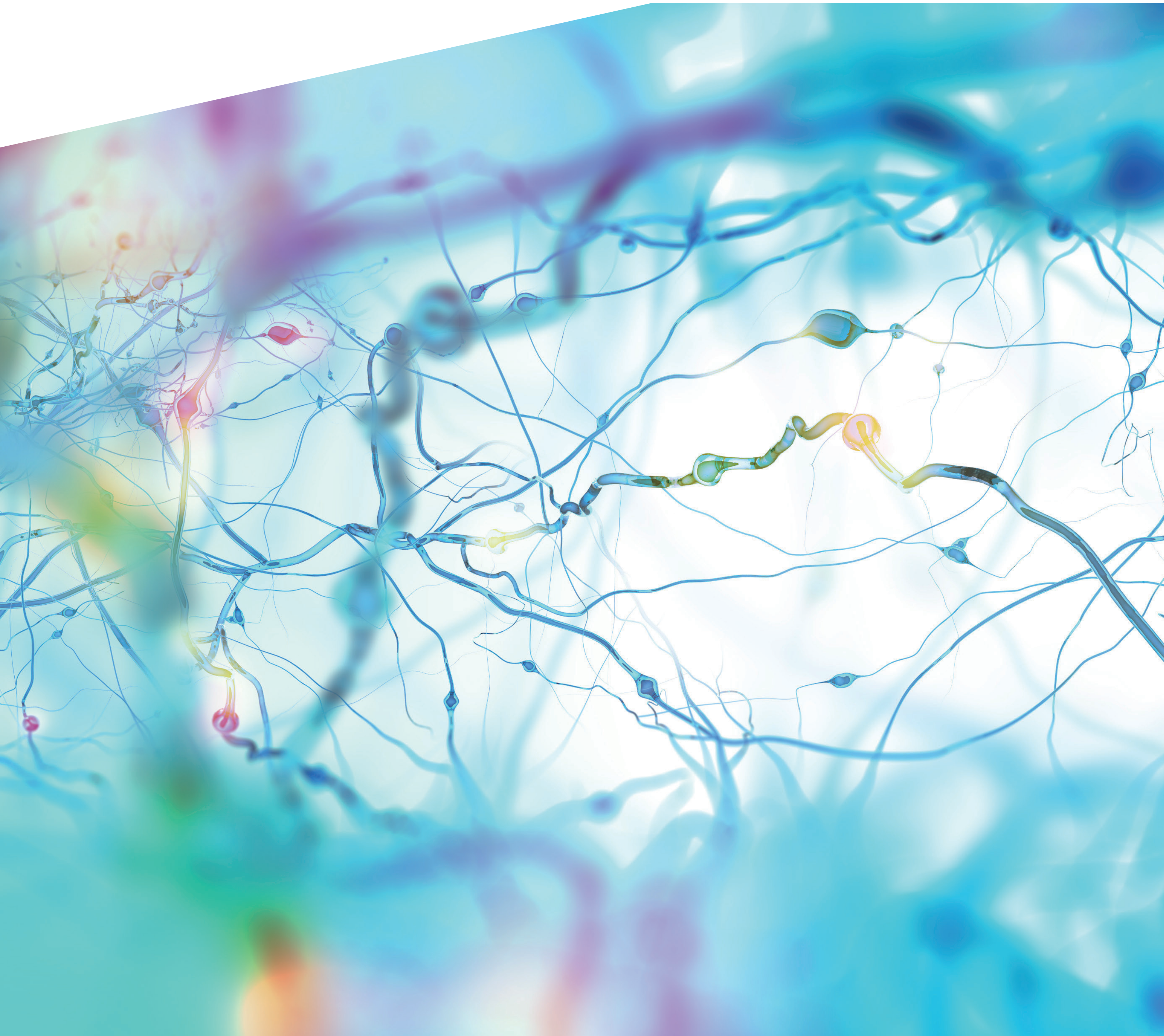
eISSN 2234-2591

Ewha Medical Journal

Vol. 48, No. 3, 2025



Ewha Medical Journal

eISSN 2234-2591

Vol. 48, No. 3, July 2025

Aims & scope

Aims

Ewha Medical Journal aims to provide medical professionals with essential healthcare information and fundamental medical knowledge. The journal will contribute to improving and serving human society based on the Christian values of education, truth, goodness, and beauty. Additionally, the journal strives to nurture young editors, enabling them to demonstrate exceptional women's editorial leadership and provide innovative learning methods.

Scope

Its scope includes:

- Up-to-date medical knowledge and skills essential for patient care
- Preparing for the future of medicine
- Effective interprofessional communication
- Ensuring gender equity and diversity
- Medical education materials
- Sharing data and protocols

Regional scope

The journal primarily focuses on Korea but welcomes submissions from researchers worldwide.

Copyright & Open access policy

Copyright and owner

Authors must declare that their work is original and that copyright is not breached. Copyright for all published material is owned by Ewha Womans University College of Medicine. Each author must sign the authorship responsibility and copyright transfer agreement, attesting to authorship criteria. The corresponding author submits the Copyright Transfer Form during submission. Authors must obtain and provide written permission for any previously published material. Submitted material will not be returned unless requested.

Open access license

EMJ is an open access journal. Articles are distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits noncommercial use, distribution, and reproduction in any medium, provided the original work is properly cited. Authors do not need permission to use tables or figures published in EMJ in other periodicals, books, or media for noncommercial purposes. For any commercial use of material from this open access journal, permission must be obtained from Ewha Womans University College of Medicine at https://www.e-emj.org/about/contact.php.

Publisher Ewha Womans University College of Medicine
Editor-in-Chief Sun Huh

Published by Ewha Womans University College of Medicine and Ewha Medical Research Institute
25, Magokdong-ro 2-gil, Gangseo-gu, Seoul 07804, Korea
Tel: +82-2-6986-6305, Fax: +82-2-6986-7026, E-mail: editor@e-emj.org

Editorial office Ewha Medical Research Institute
25, Magokdong-ro 2-gil, Gangseo-gu, Seoul 07804, Korea
Tel: +82-2-6986-6305, Fax: +82-2-6986-7026, E-mail: editor@e-emj.org

Printing correspondence by M2PI
#805, 26 Sangwon 1-gil, Seongdong-gu, Seoul 04779, Korea
Tel: +82-2-6966-4930, Fax: +82-2-6966-4945, E-mail: support@m2-pi.com

· The subscription is free. For inquiry, lease contact the editorial office (Tel: +82-2-6986-6305, E-mail: editor@e-emj.org). All the contents are also available at the EMJ website (https://www.e-emj.org).

Editorial board

eISSN 2234-2591

Ewha Medical Journal

Contents

Vol. 48, No. 3, July 2025

Editorials

Reviews

Original articles

Health statistics

Guidelines

eISSN 2234-2591

Ewha Medical Journal

Contents

Vol. 48, No. 3, July 2025

eISSN: 2234-2591
Open Access

Editorial

Ewha Med J 2025;48(3):e38
https://doi.org/10.12771/emj.2025.00710

The Ewha Medical Journal

Ewha Medical Journal's inclusion in PubMed Central and PubMed, and artificial intelligence and guidelines in this issue

Sun Huh*

Institute of Medical Education, Hallym University College of Medicine, Chuncheon, Korea

Inclusion in PubMed Central

I was delighted to receive an email from PubMed Central (PMC) on the morning of August 25, 2025, stating, "I am pleased to report this journal is now live in PMC. https://pmc.ncbi.nlm.nih.gov/journals/?term = 101606065."

Ewha Medical Journal (EMJ) applied to PMC on August 21, 2024 (Eastern Standard Time). On January 31, 2025, a scientific quality review was completed (Fig. 1) [1]. I am grateful to the PMC staff for their generous handling of EMJ and their ongoing communication regarding technical and administrative matters. Having produced Journal Article Tag Suite (JATS) XML (PMC XML) for an article in 2005, I am aware of the challenges involved in reviewing the technical quality of JATS XML. I previously applied this format to the *Journal of Educational Evaluation for Health Professions* [2]. Before 2014, all medical journals from Korea could be included in PMC if the journal was published in English and full-text XML (PMC XML) was produced. As a result, many medical journal editors in Korea transitioned their journals to English and sought PMC inclusion. However, since 2014, the criteria for scientific quality review have become stricter, causing some journals to fail in their attempts to enter PMC. EMJ was also unsuccessful on March 12, 2022. It is therefore fortunate that EMJ has now achieved inclusion in PMC after meeting both scientific and technical quality standards on its second application. The United States National Library of Medicine, which maintains PMC and PubMed, demonstrates global leadership in biomedical science for the betterment of human health. Without PMC, many local medical journals in Korea would be unable to reach the international stage, as inclusion in MEDLINE remains highly challenging for local journals. Therefore, the inclusion of abstracts in PubMed could not be anticipated.

What does it mean to be a PMC journal? When a journal is indexed in PMC, its abstracts are automatically ingested into PubMed, making all PMC-listed titles retrievable through PubMed, the world's largest biomedical literature database [3]. Inclusion in both PMC and PubMed significantly improves a journal's international visibility and often leads to increased submissions from researchers outside Korea [4]. Indeed, Korean journals typically experience a substantial rise in citation rates—sometimes as much as tenfold—after achieving PubMed and PMC indexing [5]. We expect that EMJ will enjoy a similar increase in citations once it is indexed as well.

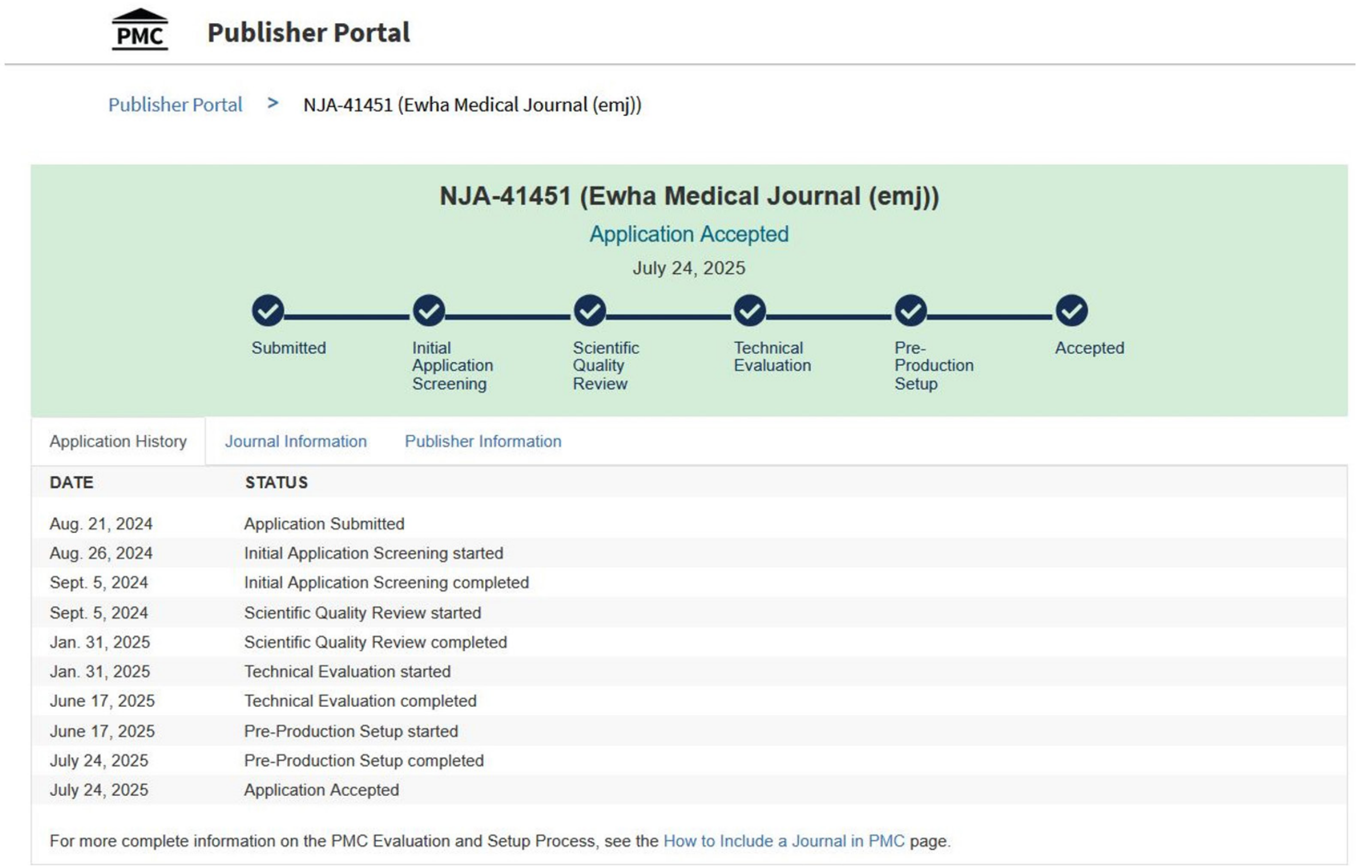
When I assumed the editorship of EMJ, one of my objectives was to have the journal included in the Directory of Open Access Journals, MEDLINE, PMC, and Scopus [6]. Of these 4, I have succeeded in 2 databases: the Directory of Open Access Journals and PMC [1]. We will soon prepare an application to MEDLINE and add previous articles to the PMC retrospective collection, extending back to 2011 [7].

Invitation of an AI article editor due to the increase in those articles

In May 2025, I invited Dr. Dohyung Rim to join the editorial board as an artificial intelligence (AI) article editor. Although I have learned about machine learning, deep learning, and large language models, I still find these topics challenging to fully understand and edit. Therefore, a specialist was appointed to handle such manuscripts. In his editorial as AI editor, Dr. Rim presented

*Corresponding email: shuh@hallym.ac.kr

Received: July 28, 2025 Accepted: July 29, 2025



PMC Publisher Portal

Publisher Portal > NJA-41451 (Ewha Medical Journal (emj))

NJA-41451 (Ewha Medical Journal (emj))

Application Accepted

July 24, 2025

DATE	STATUS
Aug. 21, 2024	Application Submitted
Aug. 26, 2024	Initial Application Screening started
Sept. 5, 2024	Initial Application Screening completed
Sept. 5, 2024	Scientific Quality Review started
Jan. 31, 2025	Scientific Quality Review completed
Jan. 31, 2025	Technical Evaluation started
June 17, 2025	Technical Evaluation completed
June 17, 2025	Pre-Production Setup started
July 24, 2025	Pre-Production Setup completed
July 24, 2025	Application Accepted

For more complete information on the PMC Evaluation and Setup Process, see the How to Include a Journal in PMC page.

Fig. 1. Application history and review results of *Ewha Medical Journal* presented in PMC Publisher Portal from August 21, 2024, to July 24, 2025.

10 guidelines for contributors to medical artificial intelligence research [8]. These guidelines will be valuable for researchers preparing machine learning and deep learning manuscripts.

In this issue, 2 papers on machine learning and deep learning models are featured. Suh et al. [9] trained trichoscopic images to detect androgenetic alopecia at an early stage. A ResNet-18 convolutional neural network, pretrained on ImageNet, was used for training. This model demonstrated high accuracy and generalizability in detecting early-stage androgenetic alopecia from trichoscopic images. The first author, Min Jung Suh, is a junior medical student who responded to reviewer comments by revising both the Python code and the manuscript with excellence. I believe this was possible because of Ewha Woman's University College of Medicine's extracurricular programs, such as the Green Ribbon Project [10], where faculty members teach students Python and deep learning. It was remarkable to see a junior student demonstrate such competency in problem-solving with a deep neural network.

The machine learning article from Statistics Korea was also noteworthy. In Korea, cause of death statistics is based on the physician's death certificate. However, direct copying of these certificates to the cause-of-death database is not possible because the initial draft by physicians must be modified in accordance with the 10th revision of the International Classification of Diseases. To minimize discrepancies between the initial draft and the final coding of cause of death, a machine learning model was developed and applied to death certificates. Among 306,898 certificates from 2022, the final cause model achieved an accuracy of 62.65%. Given that there are approximately 18,000 categories for cause of death, this result is excellent. Nevertheless, to further improve accuracy, the quality of death certificates, which serve as the foundational data source, should be enhanced [11].

Guidelines for researchers, physicians, and the general people

The reporting guidelines of CONSORT 2025 for randomized controlled studies [12], the transparent reporting of a multivariable model for individual prognosis or diagnosis (TRIPOD)-AI statement for deep learning studies in the medical field [13], and the TRIPOD-AI large language model (LLM) statement for large language model studies in medical research have been translated into Korean to facilitate easier and faster understanding among Korean researchers [14]. This translation was made possible through the support of the original guideline authors, Korean proofreading by Yoon Joo Seo, an expert manuscript editor in Korea with a background in Korean language and literature, and back-translation by Jeong-Ju Yoo, Professor of Gastroenterology at Soonchunhyang University Bucheon Hospital. For a bilingual Korean/English journal, providing such essential tools for article writing is invaluable.

The "Ten guidelines for a healthy life: Korean Medical Association Statement (2017)" was a significant achievement by the Korean Medical Association. While each recommendation was already widely known among the public, framing them within a scientifically grounded structure and supplementing them with actionable strategies added meaningful value. Additionally, the original text was expanded to include a detailed action plan. These 10 guidelines will benefit not only physicians advising clients and patients on healthy habits but also the general public interested in adopting healthier behaviors [15]. The publication of the abridged version was made possible with the cooperation of the academic leaders of the Korean Medical Association.

Finally, I would like to highlight an ecological study conducted by Dr. Eunhee Ha's lab [16]. Dr. Ha is an eminent medical researcher in environmental medicine, particularly renowned for her work on particulate matter. As noted in the "Ten guidelines for a healthy life" [15], particulate matter poses a serious health risk. In the current ecology article, the authors observed that "In winter, strong associations were observed between O_3, NO_2, and all disease outcomes. In spring, $PM_{2.5}$ and PM_{10} were strongly linked to cardiac and stroke-related visits. This connection became more pronounced in autumn, especially for NO_2 and cardiac arrest." Accordingly, targeted control of O_3 and NO_2 is an urgent priority, especially in urban environments.

One of my colleague editors of *Women's Health Nursing*, Dr. Sue Kim, recently remarked that EMJ covers a diverse range of interesting topics. My intention has always been to make EMJ a stimulating journal and a forum for communication among researchers in the health professions, including physicians, dentists, nurses, and dietitians. I am unsure how much the journal has achieved this goal, but I hope readers of EMJ find the articles in this issue enjoyable and gain at least a small sense of happiness from them.

ORCID

Sun Huh: https://orcid.org/0000-0002-8559-8640

Authors' contributions

All work was done by Sun Huh.

Conflict of interest

Sun Huh has edited *Ewha Medical Journal* since September 2023. However, he was not involved in the peer review process or decision-making. Otherwise, no potential conflict of interest relevant to this article was reported.

Funding

None.

Data availability

Not applicable.

Acknowledgments

None.

Supplementary materials

None.

References

1. Huh S. Ewha Medical Journal passed the scientific evaluation by PubMed Central and succeeded in being included in DOAJ, but failed to be accepted by Scopus. Ewha Med J 2025;48:e21. https://doi.org/10.12771/emj.2025.00024
2. Huh S. PubMed Central as a platform for the survival of open-access biomedical society journals published in Korea. Sci Ed 2021;8:153-158. https://doi.org/10.6087/kcse.247
3. Huh S. Congratulations on Child Health Nursing Research becoming a PubMed Central journal and reflections on its significance. Child Health Nurs Res 2022;28:1-4. https://doi.org/10.4094/chnr.2022.28.1.1
4. Huh S. Marking the inclusion of the Korean Journal of Women Health Nursing in PubMed Central and strategies to be promoted to a top-tier journal in the nursing category. Korean J Women Health Nurs 2022;28:165-168. https://doi.org/10.4069/kjwhn.2022.08.19
5. Jeong GH, Huh S. Increase in frequency of citation by SCIE journals of non-Medline journals after listing in an open access full-text database. Sci Ed 2014;1:24-26. https://doi.org/10.6087/kcse.2014.1.24
6. Huh S. Mission and goals of the new editor of the Ewha Medical Journal. Ewha Med J 2023;46:e9. https://doi.org/10.12771/emj.2023.e9
7. Huh S. Why do editors of local nursing society journals strive to

have their journals included in MEDLINE?: a case study of the Korean Journal of Women Health Nursing. Korean J Women Health Nurs 2023;29:147-150. https://doi.org/10.4069/kjwhn.2023.09.11.01

8. Rim D. Ten guidelines for contributors to medical artificial intelligence research. Ewha Med J 2025;48:e39. https://doi.org/10.12771/emj.2025.00717
9. Suh MJ, Ahn S, Byun JY. Automated early detection of androgenetic alopecia using deep learning on trichoscopic images from a Korean cohort: a retrospective model development and validation study. Ewha Med J 2025;48:e44. https://doi.org/10.12771/emj.2025.00486
10. Ha E. Reflections on 25 hours a day at Ewha Womans University College of Medicine from August 2021 to January 2025: a dean's farewell message. Ewha Med J 2025;48:e20. https://doi.org/10.12771/emj.2025.00045
11. Lee S, Im G. Machine learning for automated cause-of-death classification from 2021 to 2022 in Korea: development and validation of an ICD-10 prediction model. Ewha Med J 2025;48:e45. https://doi.org/10.12771/emj.2025.00675
12. Hopewell S, Chan AW, Collins GS, Hrobjartsson A, Moher D, Schulz KF, Tunn R, Aggarwal R, Berkwits M, Berlin JA, Bhandari N, Butcher NJ, Campbell MK, Chidebe RC, Elbourne D, Farmer A, Fergusson DA, Golub RM, Goodman SN, Hoffmann TC, Ioannidis JPA, Kahan BC, Knowles RL, Lamb SE, Lewis S, Loder E, Offringa M, Ravaud P, Richards DP, Rockhold FW, Schriger DL, Siegried NL, Staniszewska S, Taylor RS, Thabane L, Torgerson D, Vohra S, White IR, Boutron I. CONSORT 2025 statement: updated guideline for reporting randomized trials: a Korean translation. Ewha Med J 2025;48:e50. https://doi.org/10.12771/emj.2025.00409
13. Collins GS, Moons KG, Dhiman P, Riley RD, Beam AL, Calster BV, Ghassemi M, Liu X, Reitsma JB, Smeden MV, Boulesteix AL, Camaradou JC, Celi LA, Denaxas S, Denniston AK, Glocker B, Golub RM, Harvey H, Heinze G, Hoffman MM, Kengne AP, Lam E, Lee N, Loder EW, Maier-Hein L, Mateen BA, McCradden MD, Oakden-Rayner L, Ordish J, Parnell R, Rose S, Singh K, Wynants L, Logullo P. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods: a Korean translation. Ewha Med J 2025;48:e48. https://doi.org/10.12771/emj.2025.00668
14. Gallifant J, Afshar M, Ameen S, Aphinyanaphongs Y, Chen S, Cacciamani G, Demner-Fushman D, Dligach D, Daneshjou R, Fernandes C, Hansen LH, Landman A, Lehmann L, McCoy LG, Miller T, Moreno A, Munch N, Restrepo D, Savova G, Umeton R, Gichoya JW, Collins GS, Moons KGM, Celi LA, Bitterman DS. The TRIPOD-LLM reporting guideline for studies using large language models: a Korean translation. Ewha Med J 2025;48:e49. https://doi.org/10.12771/emj.2025.00661
15. Ahn CM, Chae JH, Choi JS, Chong YP, Chun BC, Chun EM, Kang BS, Kim DJ, Kim Y, Kwon JS, Lee SH, Lee WC, Lee YJ, Leem JH, Lim S, Park S, Shin D, Yim HW, Yoo KH, Yoon DH, Yoon HJ. Ten guidelines for a healthy life: Korean Medical Association Statement (2017). Ewha Med J 2025;48:e47. https://doi.org/10.12771/emj.2025.00696
16. Wang S, Jeong S, Ha E. Spatiotemporal associations between air pollution and emergency room visits for cardiovascular and cerebrovascular diseases in Korea using a multivariate graph autoencoder modeling approach: an ecological study. Ewha Med J 2025;48:e43. https://doi.org/10.12771/emj.2025.00640

eISSN: 2234-2591
Open Access

Editorial

Ewha Med J 2025;48(3):e39
https://doi.org/10.12771/emj.2025.00717

The Ewha Medical Journal

Ten guidelines for contributors to medical artificial intelligence research

Dohyoung Rim*

Rowan Co., Seoul, Korea

To all esteemed readers and contributors of the *Ewha Medical Journal*,

The number of medical papers incorporating deep learning technologies has increased significantly in recent years. As a result, medical journals now require specialized editorial expertise to properly evaluate these developments. Beyond validating the medical value, efficacy, and research outcomes, it has become necessary to verify the technical aspects of deep learning research methods and the analysis of results. *Ewha Medical Journal* (EMJ) is a highly progressive publication that actively supports and encourages medical research utilizing artificial intelligence (AI). Recognizing the importance of this specialized AI role, we have established the position of AI Editor, which is a role that remains rare globally. I joined EMJ this year as the AI article editor and have recently reviewed and guided several submissions. This letter aims to provide a general guide, from an engineering perspective, for preparing medical research papers that incorporate artificial intelligence. It is designed to be easily understandable, though it is neither exhaustive nor intended as a master-level resource, nor does it function as a checklist. For a reliable, comprehensive guideline, we recommend consulting the TRIPOD (transparent reporting of a multivariable model for individual prognosis or diagnosis)-AI statement [1,2].

Manuscript format: imbue the medical manuscript with engineering depth

Medical and engineering papers differ markedly in their formats, with notable distinctions in abstract structure (structured versus unstructured) and the organization of chapters. From an engineering standpoint, the overall format of an AI-driven research paper in medicine is not the primary concern; adherence to the existing format of medical journals is generally sufficient. However, certain elements deemed essential from an engineering perspective must be thoroughly addressed in the manuscript.

Reproducibility: the lifeblood of experimentation, demonstrate it transparently

Medical research employing artificial intelligence typically involves experiments using AI models, from which insights are derived through the analysis of experimental results. It is essential that readers are able to reproduce both the experiments and their outcomes based solely on the information published in the paper, even when the code is not provided. To ensure this, unlike in conventional medical papers, the following details must be clearly specified, and a dedicated section may be included for this purpose: data, preprocessing methods, model details, training details, training results, and result analysis.

Data: the foundation of research, describe it accurately and in detail

All data used in the research must be described in detail, regardless of its origin or type. When utilizing publicly available datasets, the means and methods of acquisition must be explicitly stated. For self-collected data, an even more meticulous description is

*Corresponding email: dh-rim@hanmail.net
Received: July 30, 2025 Accepted: July 30, 2025

necessary. Wherever feasible, self-collected datasets should be made publicly accessible; if this is not possible, they should at least be made available to reviewers for evaluation.

Preprocessing: a core determinant of outcomes, disclose it transparently

Data preprocessing methods have a profound impact on the results and the validity of the experiment. While general processes such as normalization and the handling of missing values are important, the way training, validation, and test datasets are composed is especially critical for the development and validation of AI models. Including any portion of the training data in the test dataset constitutes "data leakage" and renders the experimental results meaningless. Furthermore, preprocessing procedures for all 3 datasets, as well as for any required external test data, must be applied consistently to ensure the results are meaningful.

Model details: the core of the design, describe it for reproducibility

Ideally, the experimental code should be made publicly available. However, if this is not possible, the information provided in the paper should be detailed enough for readers to reconstruct the code, acquire and preprocess the data as described, and reproduce the research results. As with the data, if public disclosure of the code is not feasible, it should at least be accessible to reviewers.

Training details: the importance of environment and settings, record them meticulously

Training details encompass specifics about the environment and hyperparameters used during model training. This includes comprehensive information such as the operating system, central processing unit, graphics processing unit, Python and library versions, as well as detailed hyperparameter settings like learning rate, dropout ratio, optimizer, and its configurations. Even with identical data, preprocessing steps, and model code, the final training results and performance can vary considerably based on these settings.

Training results: visualize the process, present the results with clear metrics

Machine learning, and particularly deep learning, requires a structured training process. During training, loss and performance metrics must be monitored to determine whether the model has been adequately trained or if overfitting has occurred. Typically, these metrics are presented graphically to illustrate their progression over time. Including loss and metric graphs will greatly aid readers in understanding the training process and the outcomes achieved.

Result analysis: task-appropriate metrics, interpret the findings in depth

AI tasks in medical research broadly encompass regression and classification, as well as segmentation and generation. For regression and classification tasks, essential metrics include mean square error, receiver operating characteristic (ROC) curves, and the area under the ROC curve. For classification tasks, the confusion matrix, precision, sensitivity (also referred to as recall), and F1-score should be presented and discussed.

Comparative models: prove research value and ensure persuasiveness

If the research centers on the model itself, its superiority should be demonstrated by comparison with other models. However, in AI-powered medical research, the primary concern is the utility of the model in a medical context, rather than innovation for its own sake. Consequently, direct quantitative comparisons with other studies may be omitted when standardized experimental conditions cannot be assured. Nevertheless, even in such cases, the experimental results should be presented with sufficient persuasiveness to establish the research's value.

Discussion: bridge with clinical practice, clearly convey value

The ultimate aim of research utilizing artificial intelligence is to harness its potential as a valuable tool in clinical practice. Although the research methods and results may be more aligned with engineering content, it is vital that clinicians and medical researchers who will apply these findings can fully comprehend their significance. Therefore, the Discussion section should articulate the findings and limitations in the context of clinical practice, clearly highlighting the improvements and significance this study brings to the field. By doing so, we bridge the gap between engineering and medicine, effectively conveying the practical value of the research to readers.

Indeed, these technical details may seem somewhat unfamiliar to readers accustomed to traditional medical journal articles. This is because the writing of AI-based research papers demands far more detailed methodological explanations than traditional medical papers. However, by faithfully incorporating the elements emphasized above, researchers can substantially enhance the reliability and academic value of their work. Furthermore, such transparent reporting increases the likelihood that research findings will be applied in clinical practice, leading to meaningful advancements.

EMJ is at the forefront of this wave of change and actively encourages the participation of medical students and early-career researchers in particular. We will continue to strive to lower the barriers to entry for AI research and to support innovative studies. We hope your valuable research achievements will be widely shared through this journal, ultimately contributing to the advancement of medicine.

Sincerely,

ORCID

Dohyoung Rim: https://orcid.org/0000-0003-2022-6333

Authors' contributions

Dohyoung Rim did all the work.

Conflict of interest

Dohyoung Rim has edited *Ewha Medical Journal* since May 2025 as an AI article editor. However, he was not involved in the peer review process or decision-making. Otherwise, no potential conflict of interest relevant to this article was reported.

Funding

None.

Data availability

None.

Acknowledgments

None.

Supplementary materials

None.

References

1. Collins GS, Moons KG, Dhiman P, Riley RD, Beam AL, Van Calster B, Ghassemi M, Liu X, Reitsma JB, van Smeden M, Boulesteix AL, Camaradou JC, Celi LA, Denaxas S, Denniston AK, Glocker B, Golub RM, Harvey H, Heinze G, Hoffman MM, Kengne AP, Lam E, Lee N, Loder EW, Maier-Hein L, Mateen BA, McCradden MD, Oakden-Rayner L, Ordish J, Parnell R, Rose S, Singh K, Wynants L, Logullo P. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024;385:e078378. https://doi.org/10.1136/bmj-2023-078378
2. Collins GS, Moons KG, Dhiman P, Riley RD, Beam AL, Calster BV, Ghassemi M, Liu X, Reitsma JB, Smeden MV, Boulesteix AL, Camaradou JC, Celi LA, Denaxas S, Denniston AK, Glocker B, Golub RM, Harvey H, Heinze G, Hoffman MM, Kengne AP, Lam E, Lee N, Loder EW, Maier-Hein L, Mateen BA, McCradden MD, Oakden-Rayner L, Ordish J, Parnell R, Rose S, Singh K, Wynants L, Logullo P. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods: a Korean translation. Ewha Med J 2025;48:e48. https://doi.org/10.12771/emj.2025.00668

eISSN: 2234-2591
Open Access

Review

Ewha Med J 2025;48(3):e40
https://doi.org/10.12771/emj.2025.00703

The Ewha Medical Journal

Non-operative management of locally advanced rectal cancer with an emphasis on outcomes and quality of life: a narrative review

In Ja Park*

Department of Colon and Rectal Surgery, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea

Non-operative management, particularly the watch and wait (WW) strategy, has emerged as an alternative to total mesorectal excision for selected patients with locally advanced rectal cancer who achieve a clinical complete response (cCR) after neoadjuvant treatment. This narrative review examines oncologic outcomes, functional and quality-of-life benefits, diagnostic challenges, and surveillance requirements associated with WW compared to radical surgery. Evidence from randomized trials and international registries indicates that WW provides overall and disease-free survival rates comparable to those of surgery, provided that stringent selection criteria and intensive surveillance are maintained for 3 to 5 years. Local regrowth occurs in 15%–40% of patients—most commonly within 24 months—but salvage surgery is curative in over 90% of cases and restores oncologic equivalence. Nevertheless, distant metastasis is more frequent in patients who experience regrowth, underscoring the importance of early detection and the need for optimized systemic therapy. Accurate determination of cCR remains the primary limitation; digital rectal examination, high-resolution magnetic resonance imaging, and endoscopy, even when combined, cannot reliably exclude microscopic residual disease. Total neoadjuvant therapy increases cCR rates to 30%–60% and expands the pool of WW candidates, but also intensifies the need for standardized response definitions and surveillance algorithms. WW offers organ preservation and quality-of-life improvements without compromising survival in carefully selected patients, provided that multidisciplinary teams ensure rigorous response assessment and lifelong monitoring. Future advances in imaging, molecular biomarkers, and individualized risk stratification are expected to further enhance the safety of WW and expand eligibility to a broader patient population.

Keywords: Biomarkers; Disease-free survival; Neoadjuvant therapy; Rectal neoplasms; Quality of life

Introduction

Background

The management of locally advanced rectal cancer (LARC) has evolved markedly in recent decades, shifting from predominantly aggressive surgical interventions to more integrated, multidisciplinary approaches. Neoadjuvant chemoradiotherapy (nCRT), followed by total mesorectal excision (TME) and adjuvant chemotherapy, has become the standard treatment for LARC [1,2]. This approach has substantially improved local disease control and increased rates of sphincter preservation compared to surgery alone. However, radical surgery for rectal cancer can severely impair functional outcomes and is often associated with diminished quality of life [2]. Moreover, despite advances in local tumor control, these strategies have not consistently yielded improvements in overall survival, prompting a reassessment of treatment intensity and its attendant morbidity [1-4].

However, these protocols have resulted in a subset of patients exhibiting exceptionally favorable responses, including those classified as complete responders according to treatment criteria. Patients in this category who have undergone non-operative management have been reported to experience oncologic outcomes comparable to those who received radical resection [5,6]. This context has fostered the emergence of organ-preserving strategies, most notably the watch and wait (WW) approach, which has quickly gained prominence and has become a central focus in rec-

*Corresponding email: ipark@amc.seoul.kr

Received: June 25, 2025 Revised: July 25, 2025 Accepted: July 28, 2025

tal cancer management [5-8]. Nonetheless, significant limitations persist in the broader application of these strategies, particularly regarding the lack of reliable criteria for assessing treatment response and standardized surveillance protocols.

Objectives

The present review evaluates the treatment outcomes of non-operative management, discusses current efforts to improve patient survival, and delineates the existing constraints on real-world applications.

The paradigm shift: rationale and emergence of watch and wait

The shift toward the WW strategy is fundamentally motivated by the goal of minimizing the substantial morbidity associated with radical surgery for rectal cancer, while still achieving comparable oncologic outcomes. Although TME provides effective local tumor control, it frequently results in severe and persistent side effects, including permanent colostomy, impaired bowel function (urgency, clustering, fecal incontinence), and sexual or urinary dysfunction—all of which significantly compromise patients' quality of life [2,3].

Evidence that a notable proportion of patients—approximately 10%–25%, rising to 30%–60% with the advent of total neoadjuvant therapy (TNT)—achieve a pathological complete response (pCR) has challenged the need for radical surgery in all cases [2,9,10]. The pioneering work of Dr. Angelita Habr-Gama, who first introduced the WW concept in the early 2000s, was instrumental in this shift. Her group established strict criteria for identifying clinical complete response (cCR) using comprehensive assessments, including digital rectal examination (DRE), endoscopy, and imaging [5]. Early experiences from this group demonstrated that sustained tumor-free intervals could be achieved without surgery and that salvage treatments following local regrowth were successful, providing key evidence supporting the WW approach [5].

The development of TNT has further increased the feasibility of WW. TNT, which is defined as the administration of all chemotherapy and radiotherapy prior to surgical intervention, aims to achieve earlier systemic control of micrometastatic disease, improve patient adherence, reduce toxicity, and significantly increase tumor regression and pCR rates [9-13]. Landmark studies, including the RAPIDO and PRODIGE trials, have demonstrated significantly higher pCR rates with TNT compared to conventional nCRT, thereby expanding the pool of candidates eligible for WW [12,13]. As a result, the WW strategy has evolved from a mere avoidance of surgery to a deliberate deferral of surgery, contingent on rigorous surveillance and continuous monitoring of clinical response [10].

Importantly, the WW approach is more accurately described as "deferral of surgery" rather than "no surgery." Robust salvage pathways are essential, as highlighted by the high success rates (88%–95.4%) reported for salvage surgery in cases of local tumor regrowth [10,14]. The effectiveness and availability of salvage surgery reinforce the oncologic safety of the WW strategy. Therefore, ongoing, active surveillance is integral to WW, ensuring prompt detection and management of any tumor recurrence.

Ethics statement

This is a literature-based study; therefore, neither approval by an institutional review board nor informed consent is required.

Oncologic outcomes of watch and wait: a critical evaluation

The oncologic outcomes associated with the WW strategy for LARC have shown inconsistent results. While some reports are promising, demonstrating outcomes equivalent to those of standard radical surgical treatment, others highlight areas that require caution and further investigation.

Overall survival, disease-free survival, and disease-specific survival

Many studies and meta-analyses have reported comparable overall survival (OS) and disease-free survival (DFS) rates between patients managed with WW and those undergoing radical surgery after achieving cCR or pCR [3]. A pooled analysis of the CAO/ARO/AIO-12 and OPRA trials, which included 628 patients, found similar survival outcomes between selective WW and mandatory TME in patients with cCR or near-complete response (nCR). Specifically, the 3-year DFS (76% for WW vs. 73% for TME), distant recurrence-free survival, local recurrence-free survival, and OS were all equivalent [15]. Additionally, recent cohort analyses have reported no significant differences in OS, 5-year DFS, rates of distant metastasis, or mortality between WW and surgical groups [16-18].

However, some studies advise greater caution. One retrospective analysis comparing WW patients with cCR to pCR patients who underwent surgery reported lower survival rates for the WW group (5-year OS: 73% for WW vs. 94% for pCR; DFS: 75% for WW vs. 92% for pCR) [19]. Notably, this study identified important confounding variables, such as a higher median age in the

WW group (67.2 years vs. 57.3 years) and a substantial proportion (70%) of deaths attributable to non-cancer-related causes, both of which could influence overall survival comparisons [19]. According to real-world data from the International Watch & Wait Database (IWWD), the 2-year cumulative incidence of local regrowth was 25.2%. Conditional survival analysis indicated that patients who maintained cCR for 3 years had less than a 2% risk of systemic recurrence thereafter, during a median long-term follow-up of 55.2 months [20]. Consequently, the initial 2 years following completion of nCRT are crucial for prognostication and early detection of recurrence.

Local regrowth rates and salvage management

A major concern with WW is the risk of local tumor regrowth, with reported rates ranging from 6% to 40%, depending on study design and length of follow-up [8,10,19]. For example, the IWWD documented a 2-year cumulative incidence of local regrowth at 25% [8,20]. The OPRA trial reported local regrowth rates ranging from 27% to 40%, depending on the specific neoadjuvant therapy regimen [10].

The salvage rate for regrowth following the WW strategy varies across published studies [8,16,19,20]. Nonetheless, the majority of local regrowths can be detected early through rigorous surveillance and are amenable to effective salvage surgery. Outcomes of salvage surgery are generally favorable, with high rates of curative intent. Therefore, the feasibility and success of salvage surgery are crucial in reinforcing the oncologic safety of the WW approach.

Distant metastasis rates

Despite successful local control achieved through salvage surgery, concerns remain regarding the risk of distant metastases. Multiple studies have shown higher rates of distant metastasis in patients who experience local regrowth compared to those who maintain continuous cCR (36% vs. 1%, respectively; $P < 0.001$) [19]. Indeed, local regrowth following WW is now recognized as a significant independent risk factor for subsequent distant metastases [14]. According to the IWWD, distant metastases occurred in approximately 8% of patients managed with WW [20].

These findings suggest that an undetected primary tumor left in situ until regrowth may contribute to systemic disease progression, indicating that local regrowth could serve as an early marker of aggressive tumor biology or incomplete systemic response to treatment [20].

Critical appraisal of WW strategy

While WW demonstrates favorable outcomes when proper patient selection and surveillance are applied, it is essential to acknowledge that these outcomes are highly dependent on both early detection and successful salvage surgery for local recurrence. As such, the oncologic safety of WW critically depends on 2 key factors: the prompt identification of local regrowth and the effectiveness of subsequent salvage interventions.

Additionally, the consistently higher rates of distant metastasis observed among WW patients experiencing local regrowth suggest a potential systemic risk associated with deferring surgery. This raises important questions about whether organ preservation might inadvertently compromise systemic disease control in certain patient subgroups. Currently, the use of TNT has expanded considerably, and its oncologic outcomes have been shown to be superior to standard nCRT (Tables 1, 2), which is likely to fur-

Table 1. Location of clinical stage of patients included in TNT trials for rectal cancer

Trial	Location (cm from anal verge)	Diagnosis method	Detailed indication
nCRT vs. TNT			
RAPIDO [13]	< 16	MRI	High risk on pelvic MRI (with at least one of the following criteria: cT stage cT4a or cT4b, extramural vascular invasion, cN stage cN2, involved mesorectal fascia, or enlarged lateral lymph nodes)
STELLAR [17]	10	MRI	cT stage 3–4 and/or regional lymph node (N)–positivity
PRIODIGE 23 [14]	< 15	ERUS/MRI	Stage cT3–4 (based ERUS/MRI)
Induction vs. consolidation TNT			
CAO-ARO-AIO 12 [18]	< 12	MRI	cT3 tumor less than 6 cm from the anal verge, cT3 tumor in the middle third of the rectum (≥ 6–12 cm) with extramural tumor spread into the mesorectal fat of more than 5 mm (> cT3b), cT4 tumors, or lymph node involvement, based on MRI that was mandatory.
OPRA [10]	Require complete TME	MRI	Clinical stage II (T3–4, N-) or stage III (any T, N+) based on MRI

TNT, total neoadjuvant therapy; nCRT, neoadjvuant chemoradiotherapy; MRI, magnetic resonance imaging; cT, clinical tumor; cN, clinical nodal; ERUS, endorectal ultrasound; TME, total meosrectal excision.

Table 2. Pathologic and oncologic outcomes of TNT trials for rectal cancer

Result	R0 rate	pCR	APR rate	Local recurrence	Disease-free survival	Overall survival
nCRT vs. TNT						
RAPIDO [13]	-					
TNT	-	28	35	12.1	27.8*	81.7
nCRT	-	14	40	8.1	34.6*	80.2
STELLAR [17]						
TNT	91.5	21.8	45.1	8.4	64.5	86.5
nCRT	87.8	12.3*	41.3	11.0	62.3	75.1*
PRIODIGE 23 [14]						
TNT	95	28	14.1		67.6	81.9
nCRT	94	12	14		62.5*	76.1*
Induction vs. consolication TNT						
CAO-ARO-AIO 12 [18]						-
Consolidation	92	17	28	6	73	-
Induction	90	25*	23	5	73	-
OPRA [10]	-	-	-	-		-
Consolidation	-	-	-	-	71	-
Induction	-	-	-	-	69	-

TNT, total neoadjuvant therapy; nCRT, neoadjvuant chemoradiotherapy; pCR, pathologic complete regression; APR, abdominoperineal resection.
*P<0.005.

ther accelerate the adoption of WW. Future research must therefore focus on identifying and stratifying high-risk patients who may require intensified systemic therapy or who might not be ideal candidates for WW prior to any local recurrence.

Challenges in defining cCR

A major obstacle to the widespread adoption and standardization of the WW approach is the absence of a universally accepted and highly accurate definition of cCR [10,20,21]. At present, cCR is defined by the absence of a clinically detectable tumor on DRE, endoscopy, and imaging [10] (Fig. 1). However, these methods have significant limitations in reliably identifying true complete tumor regression, particularly at the microscopic level.

Difficulties in accurate cCR assessment

The primary challenge in determining cCR is differentiating between complete tumor regression and microscopic residual disease, as well as distinguishing post-treatment fibrosis or inflammation from viable cancer cells [22,23]. While pCR—the absence of viable tumor cells in resected surgical specimens—remains the definitive gold standard, pCR can only be confirmed after surgery and thus cannot guide the initial decision to pursue WW. Thus, discrepancies between cCR and actual pCR raise considerable concerns, as inaccurate clinical assessments may lead to understaging, increased risk of local regrowth, and potentially compromised oncologic outcomes.

Currently available diagnostic modalities, such as magnetic resonance imaging (MRI) and endoscopic biopsy, lack the sensitivity to detect microscopic residual disease, especially if tumor cells persist in deeper layers of the rectal wall [23,24]. Studies have shown that even patients categorized as having achieved cCR may harbor deeper residual tumor cells, directly contributing to local regrowth and increased risk of distant metastasis [23-26]. This diagnostic gap underscores the intrinsic limitations of existing non-invasive methods and highlights the urgent need for improved techniques capable of accurately detecting microscopic residual disease. Until such advances are achieved, strict surveillance and robust salvage strategies must remain central to WW protocols.

Criteria for optimal patient selection for WW strategy

Careful patient selection is fundamental to optimizing oncologic outcomes in WW. At present, candidates generally include those with LARC who achieve an excellent clinical response to neoadjuvant therapy [20,21]. Recently, some reports have proposed simplified response criteria after nCRT for rectal cancer, introducing the concept of a transient response group characterized as slow responders or those with near-complete response [27,28]. However, caution is warranted in selecting patients from this group, as the oncologic safety of WW in slow responders has not been clearly established.

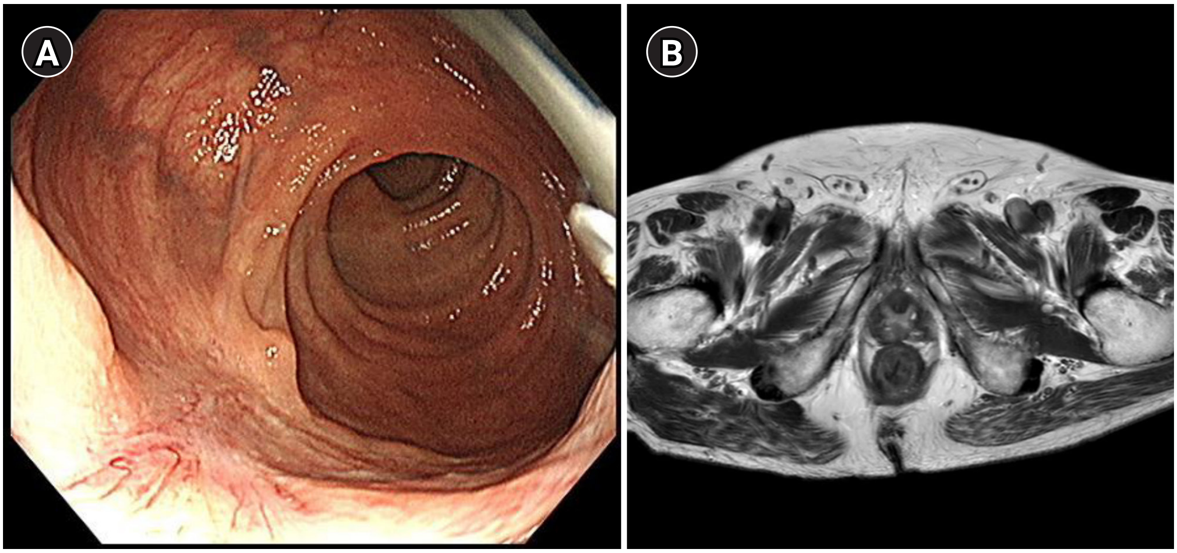


Fig. 1. Imaging diagnosis of clinical complete response after neoadjvuant therapy for rectal cancer. (A) Endoscpic image showing a whitish scar. (B) Absence of tumor signal and barely visible treatment-related sacr on magnetic resonance imaging.

Patient preferences and the ability to adhere to stringent surveillance schedules also play a critical role in patient selection [29-31]. WW is particularly appropriate for patients with low rectal tumors, where surgery would cause substantial impairment of quality of life, or for older and frail patients with significant comorbidities who are at elevated perioperative risk [31]. Thus, optimal patient selection demands a comprehensive, multidimensional assessment, balancing oncologic risks against anticipated quality-of-life gains and individual patient factors. There is a pressing need for better risk stratification tools to refine selection criteria and further personalize the WW approach. Effective implementation requires a coordinated multidisciplinary team and shared decision-making, ensuring that patients are fully informed of the potential risks and benefits.

Limitations of current diagnostic and surveillance methods for cCR

Despite substantial progress, current diagnostic and surveillance methods for determining cCR have notable limitations that hinder the consistent application and widespread standardization of WW.

DRE and endoscopy remain the cornerstone modalities for cCR assessment, enabling direct palpation and visualization of the rectal tumor bed [20,27,32]. However, these techniques are highly operator-dependent, reducing reproducibility and increasing the likelihood of missing subtle mucosal changes, submucosal residual disease, or deeper malignant foci. Endoscopic biopsy, in particular, is limited by its inability to adequately sample deeper layers, especially the muscularis propria, which leads to false-negative results [27].

MRI (1.5 T or 3 T) is the imaging method of choice for both initial staging and post-treatment assessment, providing excellent soft-tissue differentiation and visualization of mesorectal structures and lymph nodes [22,33]. Nevertheless, MRI often struggles to distinguish viable tumor from post-treatment fibrosis, edema, or inflammation, resulting in reduced accuracy for restaging after neoadjuvant therapy. The accuracy of MRI for post-treatment T-staging may be as low as 50%, and nodal staging accuracy is generally between 60% and 80% [34,35]. Additionally, variability in MRI interpretation, stemming from non-standardized reporting practices across institutions, further contributes to inconsistencies [23]. Therefore, imaging findings such as restricted diffusion or abnormal nodal morphology may suggest residual disease, but their interpretation is challenging due to significant overlap with benign post-treatment changes [27].

Diagnostic and surveillance limitations: a critical perspective

The intrinsic challenge of distinguishing fibrosis and inflammatory tissue from viable residual cancer cells on imaging—especially MRI—remains a significant diagnostic hurdle. Fibrosis induced by chemoradiotherapy often closely mimics residual disease radiologically, leading to both false-negative and false-positive assessments. This diagnostic ambiguity contributes directly to the uncertainty associated with WW outcomes and underscores the need for rigorous surveillance protocols to mitigate the risk of local regrowth. Addressing these challenges requires the development of more precise, objective imaging criteria and advanced imaging technologies capable of accurately differentiating fibrosis from viable tumor cells.

Furthermore, the high operator dependency of clinical assessments (DRE, endoscopy, endorectal ultrasound [ERUS]) and the marked variability in MRI interpretation between institutions further compromise the reproducibility and generalizability of cCR determinations. This variability undermines confidence in outcomes reported from WW registries and highlights the urgent need for standardized assessment protocols, enhanced clinician and radiologist training, and the integration of artificial intelligence (AI) and other quantitative tools to reduce inter-observer variability and improve diagnostic accuracy [36].

Establishing standardized surveillance protocols

Given the significant diagnostic challenges in defining cCR and the substantial risk of local tumor regrowth in patients managed with WW, the development of standardized and rigorously coordinated surveillance protocols is essential to ensuring patient safety and optimizing outcomes. The effectiveness of WW is fundamentally dependent on the timely detection and successful surgical management of local regrowth, underscoring the necessity for intensive and structured surveillance strategies [20,21,37].

Current surveillance practices

Early WW protocols, most notably those established by Habr-Gama, employ intensive monitoring regimens that include frequent DREs, carcinoembryonic antigen testing, endoscopic assessments, and MRI. A typical surveillance schedule involves monthly to bimonthly evaluations during the first year, quarterly evaluations in the second year, and semiannual assessments from the third year onward. Additionally, serial MRI scans are commonly performed every 6 months once an initial cCR is confirmed [6,20,37]. Many institutions implement similar protocols, generally conducting assessments every 3 months during the first 2 years and every 6 months thereafter until year 5.

The critical need for standardization

Although WW strategies are increasingly adopted worldwide, there remains considerable variability among institutions with respect to patient selection, treatment protocols, and surveillance regimens [19,20,30,38]. The National Accreditation Program for Rectal Cancer (NAPRC), for example, acknowledges the WW approach but does not provide specific clinical management or follow-up guidelines, leaving these decisions to the discretion of individual multidisciplinary teams. This lack of standardization is concerning, as inconsistent approaches can delay detection of tumor regrowth, potentially jeopardizing the outcomes of salvage surgery and increasing the risk of distant metastasis.

Given the relatively high local regrowth rates (15%–40%) [17,20,21], paired with high salvage surgery success rates (90%–95%) [37], the WW strategy should be recognized as an active management approach, not a passive observational one. Thus, the oncologic safety of WW relies less on the absolute prevention of recurrence and more on reliably detecting and effectively treating recurrence when it occurs. In this context, surveillance is not merely routine follow-up but a vital therapeutic component, transforming the paradigm from “watch and wait” to “watch, detect, and intervene promptly.”

However, there remains a gap between the ideal and the practical. While rigorous, multimodal surveillance protocols are critical for patient safety, their intensity, duration, and resource demands can be challenging for patients and healthcare systems to sustain and implement equitably [39]. The resource-intensive nature of such protocols necessitates significant expertise and training among radiologists, endoscopists, and other specialists [40]. As a result, translating the theoretical benefits of WW into standard clinical practice is often constrained by logistical barriers, patient compliance issues, and disparities in access to specialized care.

Addressing these challenges requires ongoing research aimed at optimizing surveillance strategies, striking a balance between protocol rigor, practical feasibility, patient acceptability, cost-effectiveness, and equitable access. The establishment of accredited WW centers of excellence will also be necessary to ensure quality assurance, standardized protocols, and fair patient access to specialized care pathways.

Role of multidisciplinary teams

The successful implementation of WW depends on the active participation of a highly coordinated multidisciplinary team (MDT), including surgeons, medical and radiation oncologists, radiologists, pathologists, and specialized endoscopists. The MDT must ensure accurate interpretation of clinical, radiological, and pathological findings to provide comprehensive patient assessment and ongoing monitoring. Effective shared decision-making, characterized by transparent communication regarding risks and benefits, is critical for promoting patient understanding and securing adherence to demanding surveillance protocols.

Future directions and ongoing research

The continued advancement and broader adoption of WW strategy in rectal cancer management depend significantly on dedicated research efforts. Key efforts include optimal patient selection, enhancing the accuracy of clinical response assessment, and developing effective surveillance protocols.

Current research seeks to enhance the detection of microscopic residual disease through advanced imaging technologies. Radiomics enables the extraction of detailed quantitative features from standard MRI scans, improving the prediction of cCR by correlating imaging findings with pathology and genomic data [41]. Functional MRI techniques are being investigated to better differentiate residual viable tumor from fibrosis or treatment-induced changes [42]. Novel modalities such as endorectal photoacoustic ultrasound are also under study to improve tumor response evaluation. Additionally, AI and machine learning methods are increasingly applied, utilizing deep learning algorithms to analyze imaging data and achieve more accurate assessments of treatment response [36,43].

There is an urgent need for non-invasive, highly sensitive, and specific biomarkers that can reliably predict which patients will achieve a cCR and are suitable for WW. Circulating tumor DNA, detected through liquid biopsies, has shown promise for identifying microscopic residual cancer cells and for predicting both response and recurrence [44,45]. However, current circulating biomarkers still lack sufficient specificity and clinical evidence, necessitating further validation and standardization before widespread clinical adoption.

Recent research also focuses on optimizing and intensifying TNT regimens to achieve even higher rates of complete response [46]. The OPRA trial demonstrated that long-course chemoradiotherapy followed by consolidation chemotherapy produced superior outcomes in achieving cCR and increased TME-free survival, making it a preferred approach for organ preservation [10,15]. Additional studies are exploring whether the duration and intensity of TNT regimens can be safely reduced to minimize toxicity without compromising efficacy [10-14].

Many international observational studies and clinical trials continue to validate the safety, feasibility, and efficacy of nonoperative management [16,20,47]. These studies are essential for gathering long-term outcome data and for resolving controversies regarding patient selection, cCR diagnosis, and optimal surveillance protocols. Patients undergoing WW should ideally be enrolled in prospective registries or clinical trials to contribute to this evolving evidence base. The integration of genetic profiles, molecular markers, and AI-driven predictive models represents a strong trend toward personalized medicine, enabling clinicians to tailor treatment strategies that maximize organ preservation while ensuring oncologic safety.

Conclusion

The WW strategy marks a significant evolution in the management of LARC, shifting the paradigm away from routine radical surgery and toward prioritizing organ preservation and quality of life. Its feasibility has increased with the higher rates of cCR achieved through TNT.

Despite promising survival outcomes comparable to those of standard radical surgery in patients who respond well to nCRT, important challenges persist. High rates of local regrowth remain a concern, even though salvage surgery is generally effective. Moreover, the elevated risk of distant metastasis among patients experiencing local recurrence underscores that WW is fundamentally a "deferral of surgery" rather than an outright avoidance.

Accurately identifying cCR remains a limitation, as current diagnostic tools struggle to distinguish true cCR from microscopic residual disease or post-treatment fibrosis. This complicates patient selection and mandates intensive surveillance. An alternative—careful selection of patients who may achieve a complete response during a transient observation period—has been suggested, but consensus and supporting evidence are still lacking. The lack of universally accepted and standardized surveillance protocols further restricts the widespread and equitable implementation of WW.

Future research aims to overcome these barriers through advancements in imaging technologies, AI-powered diagnostics, and the development of novel biomarkers. Ongoing clinical trials and patient registries are expected to yield essential long-term evidence to refine and optimize the WW approach.

Ultimately, the successful implementation of the WW strategy depends on a careful balance between organ preservation and oncologic safety. Rigorous patient selection, accurate diagnostic modalities, standardized surveillance protocols, and multidisciplinary shared decision-making are all essential to ensure that WW delivers optimal outcomes for patients.

ORCID

In Ja Park: https://orcid.org/0000-0001-5355-3969

Authors' contributions

All work was done by In Ja Park

Conflict of interest

No potential conflict of interest relevant to this article was reported.

Funding

None.

Data availability

Not applicable.

Acknowledgments

None.

Supplementary materials

None.

References

1. Sauer R, Liersch T, Merkel S, Fietkau R, Hohenberger W, Hess C, Becker H, Raab HR, Villanueva MT, Witzigmann H, Wittekind C, Beissbarth T, Rodel C. Preoperative versus postoperative chemoradiotherapy for locally advanced rectal cancer: results of the German CAO/ARO/AIO-94 randomized phase III trial after a median follow-up of 11 years. J Clin Oncol 2012;30:1926-1933. https://doi.org/10.1200/JCO.2011.40.1836
2. van Gijn W, Marijnen CA, Nagtegaal ID, Kranenbarg EM, Putter H, Wiggers T, Rutten HJ, Pahlman L, Glimelius B, van de Velde CJ. Preoperative radiotherapy combined with total mesorectal excision for resectable rectal cancer: 12-year follow-up of the multicentre, randomised controlled TME trial. Lancet Oncol 2011;12:575-582. https://doi.org/10.1016/S1470-2045(11)70097-3
3. Sauer R, Becker H, Hohenberger W, Rodel C, Wittekind C, Fietkau R, Martus P, Tschmelitsch J, Hager E, Hess CF, Karstens JH, Liersch T, Schmidberger H, Raab R. Preoperative versus postoperative chemoradiotherapy for rectal cancer. N Engl J Med 2004;351:1731-1740. https://doi.org/10.1056/NEJMoa040694
4. Chong CX, Koh FH, Tan HL, Sivarajah SS, Ng JL, Ho LM, Aw DK, Koo WH, Han S, Koo SL, Yip CS, Wang FQ, Foo FJ, Tan WJ. The impact of short-course total neoadjuvant therapy, long-course chemoradiotherapy, and upfront surgery on the technical difficulty of total mesorectal excision: an observational study with an intraoperative perspective. Ann Coloproctol 2024;40:451-458. https://doi.org/10.3393/ac.2023.00899.0128
5. Habr-Gama A, Perez RO, Nadalin W, Sabbaga J, Ribeiro U, Silva e Sousa AH, Campos FG, Kiss DR, Gama-Rodrigues J. Operative versus nonoperative treatment for stage 0 distal rectal cancer following chemoradiation therapy: long-term results. Ann Surg 2004;240:711-718. https://doi.org/10.1097/01.sla.0000141194.27992.32
6. Maas M, Beets-Tan RG, Lambregts DM, Lammering G, Nelemans PJ, Engelen SM, van Dam RM, Jansen RL, Sosef M, Leijtens JW, Hulsewe KW, Buijsen J, Beets GL. Wait-and-see policy for clinical complete responders after chemoradiation for rectal cancer. J Clin Oncol 2011;29:4633-4640. https://doi.org/10.1200/JCO.2011.37.7176
7. Dulskas A, Caushaj PF, Grigoravicius D, Zheng L, Fortunato R, Nunoo-Mensah JW, Samalavicius NE. International Society of University Colon and Rectal Surgeons survey of surgeons' preference on rectal cancer treatment. Ann Coloproctol 2023;39:307-314. https://doi.org/10.3393/ac.2022.00255.0036
8. van der Valk MJ, Hilling DE, Bastiaannet E, Meershoek-Klein Kranenbarg E, Beets GL, Figueiredo NL, Habr-Gama A, Perez RO, Renehan AG, van de Velde CJ. Long-term outcomes of clinical complete responders after neoadjuvant treatment for rectal cancer in the International Watch & Wait Database (IWWD): an international multicentre registry study. Lancet 2018;391:2537-2545. https://doi.org/10.1016/S0140-6736(18)31078-X
9. Patel S, Ankathi S, Haria P, Kazi M, Desouza AL, Saklani A. Impact of consolidation chemotherapy in poor responders to neoadjuvant radiation therapy: magnetic resonance imaging-based clinical-radiological correlation in high-risk rectal cancers. Ann Coloproctol 2023;39:474-483. https://doi.org/10.3393/ac.2023.00080.0011
10. Verheij FS, Omer DM, Williams H, Lin ST, Qin LX, Buckley JT, Thompson HM, Yuval JB, Kim JK, Dunne RF, Marcet J, Cataldo P, Polite B, Herzig DO, Liska D, Oommen S, Friel CM, Ternent C, Coveler AL, Hunt S, Gregory A, Varma MG, Bello BL, Carmichael JC, Krauss J, Gleisner A, Guillem JG, Temple L, Goodman KA, Segal NH, Cercek A, Yaeger R, Nash GM, Widmar M, Wei IH, Pappou EP, Weiser MR, Paty PB, Smith JJ, Wu AJ, Gollub MJ, Saltz LB, Garcia-Aguilar J. Long-term results of organ preservation in patients with rectal adenocarcinoma treated with total neoadjuvant therapy: the randomized phase II OPRA trial. J Clin Oncol 2024;42:500-506. https://doi.org/10.1200/JCO.23.01208
11. Sychev S, Ponomarenko A, Chernyshov S, Alekseev M, Mamedli Z, Kuzmichev D, Polynovskiy A, Rybakov E. Total neoadjuvant therapy in rectal cancer: a network meta-analysis of randomized trials. Ann Coloproctol 2023;39:289-300. https://doi.org/10.3393/ac.2022.00920.0131

12. Conroy T, Bosset JF, Etienne PL, Rio E, Francois E, Mesgouez-Nebout N, Vendrely V, Artignan X, Bouche O, Gargot D, Boige V, Bonichon-Lamichhane N, Louvet C, Morand C, de la Fouchardiere C, Lamfichekh N, Juzyna B, Jouffroy-Zeller C, Rullier E, Marchal F, Gourgou S, Castan F, Borg C. Neoadjuvant chemotherapy with FOLFIRINOX and preoperative chemoradiotherapy for patients with locally advanced rectal cancer (UNICANCER-PRODIGE 23): a multicentre, randomised, open-label, phase 3 trial. Lancet Oncol 2021;22:702-715. https://doi.org/10.1016/S1470-2045(21)00079-6
13. Bahadoer RR, Dijkstra EA, van Etten B, Marijnen CA, Putter H, Kranenbarg EM, Roodvoets AG, Nagtegaal ID, Beets-Tan RG, Blomqvist LK, Fokstuen T, Ten Tije AJ, Capdevila J, Hendriks MP, Edhemovic I, Cervantes A, Nilsson PJ, Glimelius B, van de Velde CJ, Hospers GA. Short-course radiotherapy followed by chemotherapy before total mesorectal excision (TME) versus preoperative chemoradiotherapy, TME, and optional adjuvant chemotherapy in locally advanced rectal cancer (RAPIDO): a randomised, open-label, phase 3 trial. Lancet Oncol 2021;22:29-42. https://doi.org/10.1016/S1470-2045(20)30555-6
14. Lin W, Wee IJ, Seow-En I, Chok AY, Tan EK. Survival outcomes of salvage surgery in the watch-and-wait approach for rectal cancer with clinical complete response after neoadjuvant chemoradiotherapy: a systematic review and meta-analysis. Ann Coloproctol 2023;39:447-456. https://doi.org/10.3393/ac.2022.01221.0174
15. Williams H, Fokas E, Diefenhardt M, Lee C, Verheij FS, Omer DM, Lin ST, Dunne RF, Marcet J, Cataldo P, Polite B, Piso P, Polat B, Dapper H, Ghadimi M, Hofheinz RD, Qin LX, Saltz LB, Wu AJ, Gollub MJ, Smith JJ, Weiser MR, Rodel C, Garcia-Aguilar J. Survival among patients treated with total mesorectal excision or selective watch-and-wait after total neoadjuvant therapy: a pooled analysis of the CAO/ARO/AIO-12 and OPRA randomized phase II trials. Ann Oncol 2025;36:543-547. https://doi.org/10.1016/j.annonc.2025.01.006
16. Renehan AG, Malcomson L, Emsley R, Gollins S, Maw A, Myint AS, Rooney PS, Susnerwala S, Blower A, Saunders MP, Wilson MS, Scott N, O'Dwyer ST. Watch-and-wait approach versus surgical resection after chemoradiotherapy for patients with rectal cancer (the OnCoRe project): a propensity-score matched cohort analysis. Lancet Oncol 2016;17:174-183. https://doi.org/10.1016/S1470-2045(15)00467-2
17. Jin J, Tang Y, Hu C, Jiang LM, Jiang J, Li N, Liu WY, Chen SL, Li S, Lu NN, Cai Y, Li YH, Zhu Y, Cheng GH, Zhang HY, Wang X, Zhu SY, Wang J, Li GF, Yang JL, Zhang K, Chi Y, Yang L, Zhou HT, Zhou AP, Zou SM, Fang H, Wang SL, Zhang HZ, Wang XS, Wei LC, Wang WL, Liu SX, Gao YH, Li YX. Multicenter, randomized, phase III trial of short-term radiotherapy plus chemotherapy versus long-term chemoradiotherapy in locally advanced rectal cancer (STELLAR). J Clin Oncol 2022;40:1681-1692. https://doi.org/10.1200/JCO.21.01667
18. Fokas E, Schlenska-Lange A, Polat B, Klautke G, Grabenbauer GG, Fietkau R, Kuhnt T, Staib L, Brunner T, Grosu AL, Kirste S, Jacobasch L, Allgauer M, Flentje M, Germer CT, Grutzmann R, Hildebrandt G, Schwarzbach M, Bechstein WO, Sulberg H, Friede T, Gaedcke J, Ghadimi M, Hofheinz RD, Rodel C. Chemoradiotherapy plus induction or consolidation chemotherapy as total neoadjuvant therapy for patients with locally advanced rectal cancer: long-term results of the CAO/ARO/AIO-12 randomized clinical trial. JAMA Oncol 2022;8:e215445. https://doi.org/10.1001/jamaoncol.2021.5445
19. Smith JJ, Strombom P, Chow OS, Roxburgh CS, Lynn P, Eaton A, Widmar M, Ganesh K, Yaeger R, Cercek A, Weiser MR, Nash GM, Guillem JG, Temple LK, Chalasani SB, Fuqua JL, Petkovska I, Wu AJ, Reyngold M, Vakiani E, Shia J, Segal NH, Smith JD, Crane C, Gollub MJ, Gonen M, Saltz LB, Garcia-Aguilar J, Paty PB. Assessment of a watch-and-wait strategy for rectal cancer in patients with a complete response after neoadjuvant therapy. JAMA Oncol 2019;5:e185896. https://doi.org/10.1001/jamaoncol.2018.5896
20. Fernandez LM, Sao Juliao GP, Figueiredo NL, Beets GL, van der Valk MJ, Bahadoer RR, Hilling DE, Meershoek-Klein Kranenbarg E, Roodvoets AG, Renehan AG, van de Velde CJ, Habr-Gama A, Perez RO. Conditional recurrence-free survival of clinical complete responders managed by watch and wait after neoadjuvant chemoradiotherapy for rectal cancer in the International Watch & Wait Database: a retrospective, international, multicentre registry study. Lancet Oncol 2021;22:43-50. https://doi.org/10.1016/S1470-2045(20)30557-X
21. Gani C, Fokas E, Polat B, Ott OJ, Diefenhardt M, Konigsrainer A, Boke S, Kirschniak A, Bachmann R, Wichmann D, Bitzer M, Clasen S, Grosse U, Hoffmann R, Gotz M, Hofheinz RD, Germer E, Germer CT, Fietkau R, Martus P, Zips D, Rodel C. Organ preservation after total neoadjuvant therapy for locally advanced rectal cancer (CAO/ARO/AIO-16): an open-label, multicentre, single-arm, phase 2 trial. Lancet Gastroenterol Hepatol 2025;10:562-572. https://doi.org/10.1016/S2468-1253(25)00049-4
22. Patel UB, Taylor F, Blomqvist L, George C, Evans H, Tekkis P, Quirke P, Sebag-Montefiore D, Moran B, Heald R, Guthrie A, Bees N, Swift I, Pennert K, Brown G. Magnetic resonance imaging-detected tumor response for locally advanced rectal cancer predicts survival outcomes: MERCURY experience. J Clin Oncol 2011;29:3753-3760. https://doi.org/10.1200/JCO.2011.

34.9068

23. Park SH, Cho SH, Choi SH, Jang JK, Kim MJ, Kim SH, Lim JS, Moon SK, Park JH, Seo N. MRI Assessment of complete response to preoperative chemoradiation therapy for rectal cancer: 2020 guide for practice from the Korean Society of Abdominal Radiology. Korean J Radiol 2020;21:812-828. https://doi.org/10.3348/kjr.2020.0483
24. van der Sande ME, Maas M, Melenhorst J, Breukink SO, van Leerdam ME, Beets GL. Predictive value of endoscopic features for a complete response after chemoradiotherapy for rectal cancer. Ann Surg 2021;274:e541-e547. https://doi.org/10.1097/SLA.0000000000003718
25. Williams H, Lee C, Garcia-Aguilar J. Nonoperative management of rectal cancer. Front Oncol 2024;14:1477510. https://doi.org/10.3389/fonc.2024.1477510
26. Wang QX, Zhang R, Xiao WW, Zhang S, Wei MB, Li YH, Chang H, Xie WH, Li LR, Ding PR, Chen G, Zeng ZF, Wang WH, Wan XB, Gao YH. The watch-and-wait strategy versus surgical resection for rectal cancer patients with a clinical complete response after neoadjuvant chemoradiotherapy. Radiat Oncol 2021;16:16. https://doi.org/10.1186/s13014-021-01746-0
27. Custers PA, Beets GL, Bach SP, Blomqvist LK, Figueiredo N, Gollub MJ, Martling A, Melenhorst J, Ortega CD, Perez RO, Smith JJ, Lambregts DM, Beets-Tan RG, Maas M. An international expert-based consensus on the definition of a clinical near-complete response after neoadjuvant (chemo)radiotherapy for rectal cancer. Dis Colon Rectum 2024;67:782-795. https://doi.org/10.1097/DCR.0000000000003209
28. Gambacorta MA, Masciocchi C, Chiloiro G, Meldolesi E, Macchia G, van Soest J, Peters F, Collette L, Gerard JP, Ngan S, Rodel CC, Damiani A, Dekker A, Valentini V. Timing to achieve the highest rate of pCR after preoperative radiochemotherapy in rectal cancer: a pooled analysis of 3085 patients from 7 randomized trials. Radiother Oncol 2021;154:154-160. https://doi.org/10.1016/j.radonc.2020.09.026
29. Son GM. Organ preservation for early rectal cancer using preoperative chemoradiotherapy. Ann Coloproctol 2023;39:191-192. https://doi.org/10.3393/ac.2023.00409.0058
30. Habr-Gama A, Lynn PB, Jorge JM, Sao Juliao GP, Proscurshim I, Gama-Rodrigues J, Fernandez LM, Perez RO. Impact of organ-preserving strategies on anorectal function in patients with distal rectal cancer following neoadjuvant chemoradiation. Dis Colon Rectum 2016;59:264-269. https://doi.org/10.1097/DCR.0000000000000543
31. Jung WB. Beyond survival: a comprehensive review of quality of life in rectal cancer patients. Ann Coloproctol 2024;40:527-537. https://doi.org/10.3393/ac.2024.00745.0106
32. Thompson HM, Omer DM, Lin S, Kim JK, Yuval JB, Verheij FS, Qin LX, Gollub MJ, Wu AJ, Lee M, Patil S, Hezel AF, Marcet JE, Cataldo PA, Polite BN, Herzig DO, Liska D, Oommen S, Friel CM, Ternent CA, Coveler AL, Hunt SR, Garcia-Aguilar J. Organ preservation and survival by clinical response grade in patients with rectal cancer treated with total neoadjuvant therapy: a secondary analysis of the OPRA randomized clinical trial. JAMA Netw Open 2024;7:e2350903. https://doi.org/10.1001/jamanetworkopen.2023.50903
33. Beets-Tan RG, Lambregts DM, Maas M, Bipat S, Barbaro B, Curvo-Semedo L, Fenlon HM, Gollub MJ, Gourtsoyianni S, Halligan S, Hoeffel C, Kim SH, Laghi A, Maier A, Rafaelsen SR, Stoker J, Taylor SA, Torkzad MR, Blomqvist L. Magnetic resonance imaging for clinical management of rectal cancer: updated recommendations from the 2016 European Society of Gastrointestinal and Abdominal Radiology (ESGAR) consensus meeting. Eur Radiol 2018;28:1465-1475. https://doi.org/10.1007/s00330-017-5026-2
34. Celik H, Barlik F, Sokmen S, Terzi C, Canda AE, Sagol O, Sarioglu S, Unlu M, Bilkay Gorken I, Arican Alicikus Z, Oztop I. Diagnostic performance of magnetic resonance imaging in preoperative local staging of rectal cancer after neoadjuvant chemoradiotherapy. Diagn Interv Radiol 2023;29:219-227. https://doi.org/10.4274/dir.2022.221333
35. Liu B, Sun C, Zhao X, Liu L, Liu S, Ma H. The value of multimodality MR in T staging evaluation after neoadjuvant therapy for rectal cancer. Technol Health Care 2024;32:615-627. https://doi.org/10.3233/THC-220798
36. Kim M, Park T, Oh BY, Kim MJ, Cho BJ, Son IT. Performance reporting design in artificial intelligence studies using image-based TNM staging and prognostic parameters in rectal cancer: a systematic review. Ann Coloproctol 2024;40:13-26. https://doi.org/10.3393/ac.2023.00892.0127
37. Habr-Gama A, Gama-Rodrigues J, Sao Juliao GP, Proscurshim I, Sabbagh C, Lynn PB, Perez RO. Local recurrence after complete clinical response and watch and wait in rectal cancer after neoadjuvant chemoradiation: impact of salvage therapy on local disease control. Int J Radiat Oncol Biol Phys 2014;88:822-828. https://doi.org/10.1016/j.ijrobp.2013.12.012
38. Meng C, Shu W, Sun L, Wu S, Wei P, Gao J, Shi J, Li Y, Yang Z, Yao H, Zhang Z. Rectal cancer approach strategies after neoadjuvant treatment: a systematic review and network meta-analysis. Int J Surg 2025;111:3078-3092. https://doi.org/10.1097/JS9.0000000000002290
39. Crean R, Glyn T, McCombie A, Frizelle F. Comparing outcomes and cost in surgery versus watch & wait surveillance of

patients with rectal cancer post neoadjuvant long course chemoradiotherapy. ANZ J Surg 2024;94:1151-1160. https://doi.org/10.1111/ans.18916

40. Feeney G, Sehgal R, Sheehan M, Hogan A, Regan M, Joyce M, Kerin M. Neoadjuvant radiotherapy for rectal cancer management. World J Gastroenterol 2019;25:4850-4869. https://doi.org/10.3748/wjg.v25.i33.4850

41. Rai J, Mai DV, Drami I, Pring ET, Gould LE, Lung PF, Glover T, Shur JD, Whitcher B, Athanasiou T, Jenkins JT. MRI radiomics prediction modelling for pathological complete response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer: a systematic review and meta-analysis. Abdom Radiol (NY) 2025;Apr 28 [Epub]. https://doi.org/10.1007/s00261-025-04953-5

42. Pham TT, Liney GP, Wong K, Barton MB. Functional MRI for quantitative treatment response prediction in locally advanced rectal cancer. Br J Radiol 2017;90:20151078. https://doi.org/10.1259/bjr.20151078

43. Boldrini L, Charles-Davies D, Romano A, Mancino M, Nacci I, Tran HE, Bono F, Boccia E, Gambacorta MA, Chiloiro G. Response prediction for neoadjuvant treatment in locally advanced rectal cancer patients-improvement in decision-making: a systematic review. Eur J Surg Oncol 2025;51:109463. https://doi.org/10.1016/j.ejso.2024.109463

44. Morais M, Pinto DM, Machado JC, Carneiro S. CtDNA on liquid biopsy for predicting response and prognosis in locally advanced rectal cancer: a systematic review. Eur J Surg Oncol 2022;48:218-227. https://doi.org/10.1016/j.ejso.2021.08.034

45. Nassar A, Aly NE, Jin Z, Aly EH. CtDNA as a predictor of outcome after curative resection for locally advanced rectal cancer: systematic review and meta-analysis. Colorectal Dis 2024;26:1346-1358. https://doi.org/10.1111/codi.17039

46. Liu J, Xu X, Zhong H, Yu M, Abuduaini N, Fingerhut A, Cai Z, Feng B. Optimizing total neoadjuvant therapy in locally advanced rectal cancer: risk stratification should not be overlooked. Future Oncol 2025;21:1951-1960. https://doi.org/10.1080/14796694.2025.2507560

47. Wang L, Zhang XY, Zhao YM, Li SJ, Li ZW, Sun YS, Wang WH, Wu AW. Intentional watch and wait or organ preservation surgery following neoadjuvant chemoradiotherapy plus consolidation CAPEOX for MRI-defined low-risk rectal cancer: findings from a prospective phase 2 Trial (PKUCH-R01 trial, NCT02860234). Ann Surg 2023;277:647-654. https://doi.org/10.1097/SLA.0000000000005507

eISSN: 2234-2591
Open Access

Review

Ewha Med J 2025;48(3):e41
https://doi.org/10.12771/emj.2025.00479

The Ewha Medical Journal

A history of 20 years of medical education at Ewha Womans University College of Medicine

Ivo Kwon[1], Somi Jeong[2*], Seung-Jung Kim[2,3], Ara Ko[2], Hyeonji Jeon[1]

[1]Department of Medical Education, Ewha Womans University College of Medicine, Seoul, Korea
[2]Ewha Medical Education Center, Ewha Womans University College of Medicine, Seoul, Korea
[3]Department of Nephrology, Ewha Womans University College of Medicine, Seoul, Korea

The study aims to examine the 20-year developmental trajectory of medical education at Ewha Womans University College of Medicine (2004–2025). It analyzes educational support documents, self-evaluation reports, and Curriculum Committee meeting minutes to illuminate both the direction and significance of Ewha's medical education reforms. Key milestones include the formal establishment of the Medical Education Office in 2004 and the subsequent founding of the Department of Medical Education in 2005. Major innovations over this period encompass the expansion of objective structured clinical examinations and the introduction of problem-based learning modules. Additional advancements include the establishment of the Ewha Medical Simulation Center and Learning Resource Center, as well as the reversion to an undergraduate medical college format in 2015. The college has also prioritized faculty development workshops and medical education seminars, implemented the Ewha Social Active Communication program, and introduced team-based learning. Noteworthy initiatives include the enhancement of student research capacity and the launch of a dedicated medical education newsletter. In 2022, the Medical Education Office was reorganized as the Ewha Center for Medical Education, marking a new era of integrated leadership and expanded educational initiatives. Ewha has consistently achieved high accreditation statuses, reflecting ongoing excellence in curriculum development, assessment, and faculty development. This progress demonstrates the dedication and collaboration of both faculty and staff, resulting in a robust educational framework. The institution's continuous growth serves not only as a testament to past achievements but also as a foundation for future advancements in Ewha's medical education, with the ultimate aim of cultivating women leaders in Korean healthcare.

Keywords: Anniversaries and special events; Communication; Leadership; Medical education; Problem-based learning

Introduction

This paper systematically examines the developmental trajectory of medical education at Ewha Womans University College of Medicine. To achieve this, educational support documents generated over the past 2 decades were collected and analyzed, with special attention paid to temporal changes and major curricular themes. The primary data sources consisted of (1) the Ewha Medical Education Office's educational support archives (2004–2025) [1], (2) the College's self-evaluation reports [2], and (3) minutes of the Curriculum Committee meetings (2004–2025) [3].

Through this analysis, the study aims to illuminate both the direction and significance of Ewha's medical education reforms. By reflecting on the past and examining the present, these findings are expected to offer critical insights for designing the future of the College's educational initiatives. Although Ewha has pursued women's medical education since its founding in 1945 and achieved numerous milestones, this investigation focuses on the last 20 years. This period was selected because the establishment of the Medical Education Office (now the Ewha Center for Medical Education) and the Department of Medical Education during this time provided the institutional framework necessary for a coordinated, scholarly approach to "medical education" as a distinct field (Supplement 1).

*Corresponding email: angelica66@ewha.ac.kr
Received: May 30, 2025 Accepted: June 20, 2025

Current status of medical education at Ewha Womans University College of Medicine

Key figures in medical education at Ewha

The formal establishment of the Medical Education Office in 2004 marked the beginning of systematic organizational management. Professor Soon Nam Lee (Internal Medicine) was appointed as the first Director, with Professor Bok Kyu (Ivo) Kwon (Medical Humanities) serving as Deputy Director and Professor Jae Jin Han (Thoracic Surgery) contributing in a joint capacity. Professor Lee subsequently expanded her leadership by serving as Dean of the College of Medicine. In 2005, the creation of the Department of Medical Education laid the scholarly foundation for Ewha's medical education efforts.

Starting in 2007, Professor Jae-Jin Han took on the directorship, with Professor Bok-Kyu Kwon continuing as Deputy Director. During this time, the faculty base broadened through the involvement of BK21 Research Professors Jung Hee An, Eun Kyung Eo (Emergency Medicine), and Sun Young Hong (Internal Medicine), each serving as additional Deputy Directors.

In 2013, Professor Hee Jung Choi (Internal Medicine) became Director (then titled "Head"), supported by Deputy Directors Professor Hye-Kyung Jung (Internal Medicine) and Professor Eun Kyung Eo. From 2014 onward, Professor Yeong Seon Hong led the office, with Professors Hye Kyung Jung and Han Soo Kim (Otorhinolaryngology) serving as Deputy Directors, further consolidating and deepening the program.

In 2018, Professor Do Sang Cho (Neurosurgery) assumed the directorship, with Deputy Directors Professor Claire Junga Kim (Medical Humanities) and Professor Hee Sung Lee (Surgery). The following year, Professor Bok-Kyu (Ivo) Kwon returned as Director, with Professors Claire Junga Kim and Hee Seong Lee continuing as Deputies.

In 2021, Professor Hyeon-Jong Yang (Parasitology) took on the role of Director, joined by Deputy Directors Professor Chung Hyun Tae (Internal Medicine) and Hee Seong Lee, thereby maintaining continuity in educational operations.

The Medical Education Office was expanded into the Ewha Center for Medical Education in 2022, with the Center Director also serving as Associate Dean for Education. Professor Wook Bum Pyun (Internal Medicine) became the first Center Director, assisted by Deputy Directors Chung Hyun Tae and Hee Seong Lee. Simultaneously, Professor So-Mi Jeong, an education specialist, joined the Center as a dedicated educational professional.

Since 2024, Professor Seung-Jung Kim (Internal Medicine) has

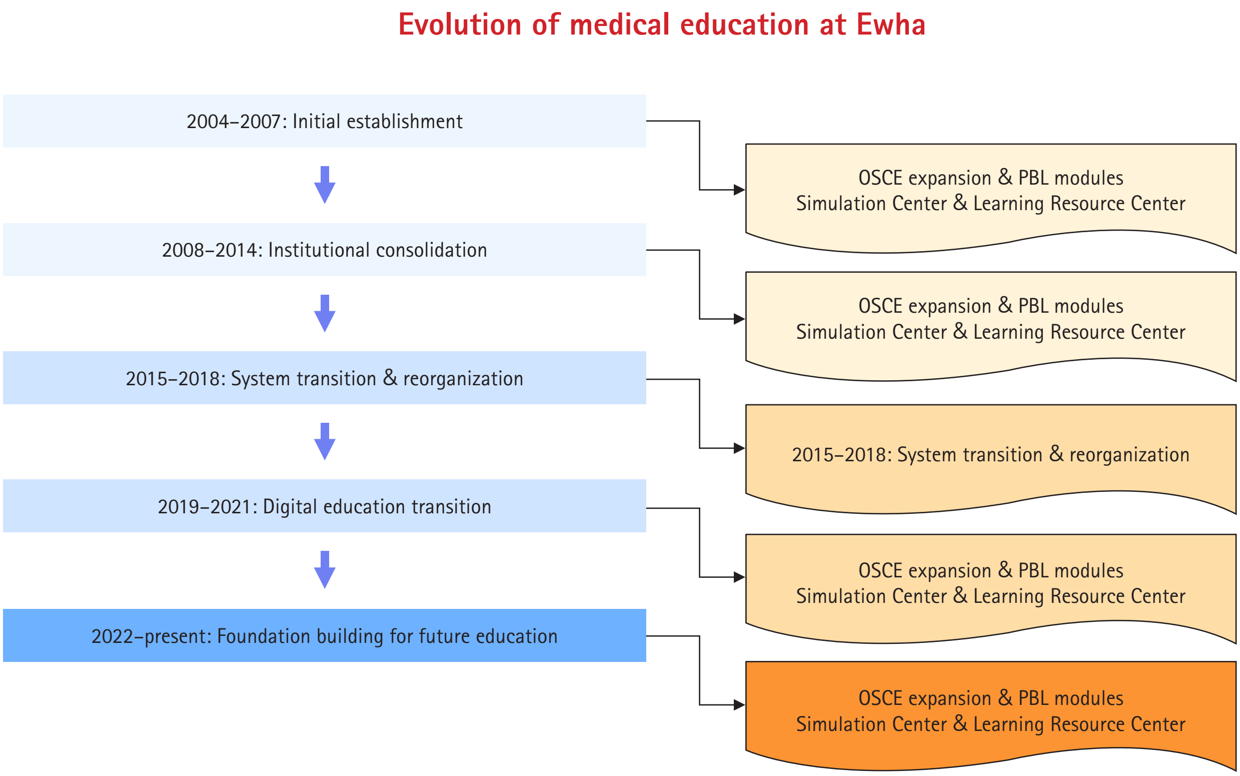

Evolution of medical education at Ewha
2004–2007: Initial establishment
2008–2014: Institutional consolidation
2015–2018: System transition & reorganization
2019–2021: Digital education transition
2022–present: Foundation building for future education
OSCE expansion & PBL modules
Simulation Center & Learning Resource Center
OSCE expansion & PBL modules
Simulation Center & Learning Resource Center
2015–2018: System transition & reorganization
OSCE expansion & PBL modules
Simulation Center & Learning Resource Center
OSCE expansion & PBL modules
Simulation Center & Learning Resource Center

Fig. 1. Evolution of medical education in Ewha (2004–2025). OSCE, objective structured clinical examination; PBL, problem-based learning.

served as Center Director and Associate Dean for Education, with Professors Chung Hyun Tae and Min Gyeong Jeong as Deputy Directors.

This succession of leadership demonstrates the progressive institutionalization and intellectual maturation of medical education at Ewha Womans University College of Medicine.

Principal phases in the evolution of medical education at Ewha

The development of medical education at Ewha Womans University College of Medicine can be divided into 5 distinct phases: (1) initial establishment (2004–2007), (2) institutional consolidation (2008–2014), (3) system transition and reorganization (2015–2018), (4) digital education transition (2019–2021), and (5) foundation-building for future education (2022–present).

The salient changes and accomplishments of each phase are detailed below (Fig. 1).

Initial establishment (2004–2007)

This period marks the formal organization of Ewha's medical education framework. With the official launch of the Medical Education Office in 2004, a structured administrative system was established. Key innovations during this time included the expansion of the objective structured clinical examination (OSCE) and the development of problem-based learning (PBL) modules. Leadership development camps, aimed at fostering female medical leaders, became an institutional tradition, and systematic curricular research was initiated in preparation for the shift to a graduate-entry medical college. The creation of the Department of Medical Education in 2005 provided a dedicated academic foundation for the field. In 2007, the opening of the Ewha Medical Simulation Center and the introduction of a Learning Resource Center (LRC) further strengthened the educational infrastructure.

Institutional consolidation (2008–2014)

During this stage, operational stability was achieved and the groundwork was laid for an eventual return to an undergraduate medical college structure. Systematic support for both student education and faculty development was established, and an outcomes-based assessment framework for graduation was put in place. Course syllabi were revised, and co-curricular programs such as Ewha Social Active Communication (ESAC) and the Standardized Patient Instructor initiative were introduced to enhance medical professionalism and social competencies. In the latter part of this phase, preparations began for developing a pre-medical track, setting the stage for future curriculum changes.

System transition and reorganization (2015–2018)

The introduction of the pre-medical program in 2015 marked Ewha's return from a graduate-entry system to a traditional medical college format. This transition prompted significant reforms in pedagogical methods, faculty involvement, and administrative operations. Comprehensive restructuring was undertaken across learning environments, assessment methodologies, and support systems. The new undergraduate curriculum was formally launched in 2017, symbolizing this pivotal change. A dedicated task force for curriculum reform was established, and important discussions began regarding the implementation of computer-based testing.

Digital education transition (2019–2021)

The relocation to the Magok campus spurred rapid expansion of digital infrastructure, including the introduction of a learning management system, an integrated portfolio platform, and centralized management of the Simulation Center. The global COVID-19 (coronavirus disease 2019) pandemic further accelerated the adoption of online instruction and remote assessments, enhancing educational flexibility and sustainability. This period was characterized by a systematic response to the demands of digital transformation in medical education.

Foundation-building for future education (2022–present)

The expansion of the Medical Education Office into the Ewha Center for Medical Education in 2022 marked the beginning of a new era. The Center Director simultaneously serves as Associate Dean for Education, reflecting integrated leadership. Key initiatives include public disclosure of feedback and course evaluations, regular medical education seminars, the revitalization of a journal club, and the publication of an internal newsletter, all of which contribute to a robust educational culture. Ongoing long-term initiatives include refinement of an outcomes-based curriculum, phasing out PBL in favor of expanded team-based learning (TBL), operating a task force for a 6-year integrated curriculum, and implementing a comprehensive system for educational quality management.

Core themes in Ewha's medical-education enterprise

Faculty development

Faculty development initiatives are organized under 2 principal strands: faculty development workshops and medical education seminars (Fig. 2, Supplement 2).

Faculty development workshops: Although instructional workshops predate 2004, they expanded considerably after the Medical

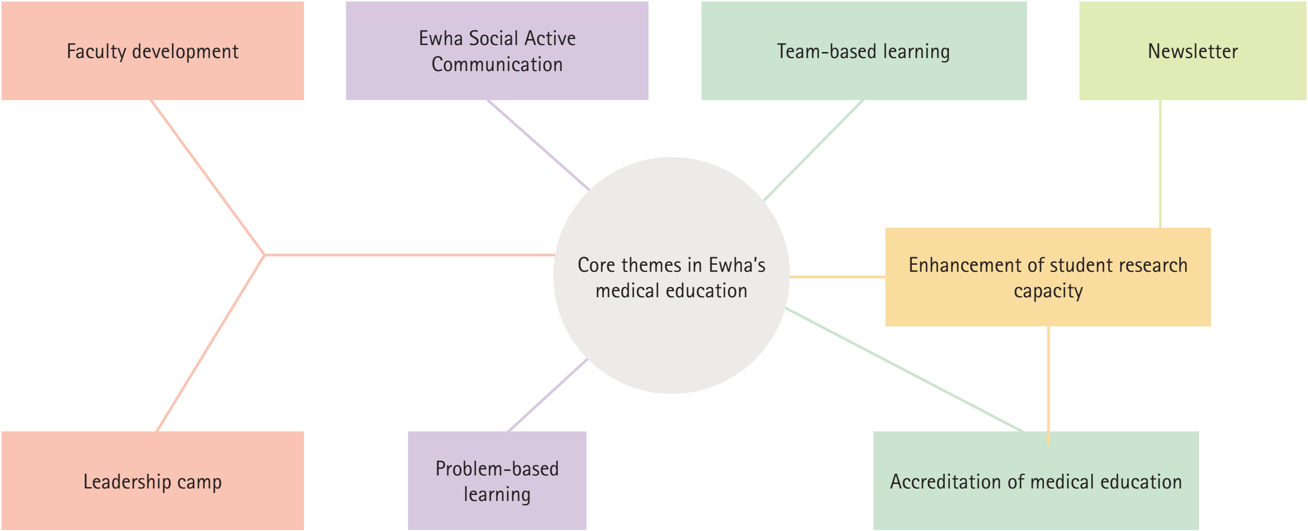

Faculty development
Ewha Social Active Communication
Team-based learning
Newsletter
Core themes in Ewha's medical education
Enhancement of student research capacity
Leadership camp
Problem-based learning
Accreditation of medical education

Fig. 2. Core themes of Ewha medical education (2025).

Education Office was established, with a practical emphasis on implementing the new curriculum. Early workshops focused on curriculum design, item writing for the national licensing examination, and clinical performance examination (CPX) assessment, providing a foundation for broad faculty participation. Beginning in 2006, specialized workshops for junior faculty were introduced, along with sessions on PBL, TBL, and integrated curriculum design. As the program evolved, topics diversified to include medical education research methods, mentoring, and department-specific pedagogical sessions. By the mid-2010s, the workshops shifted toward developing assessment literacy, featuring OSCE/CPX scoring conferences, advanced item-construction training, and course outcome workshops aligned with competency-based education. An annual college-wide retreat was established to foster a shared educational philosophy and collegial dialogue. Most workshops are now embedded in the academic calendar, with additional ad hoc sessions provided as new needs emerge.

Medical education seminars: Originally conceived as individual lectures on curriculum design and teaching methods, the seminar series broadened its scope after the 2011 "Social & Smart Learning" symposium, which highlighted new educational paradigms. In 2015, a thematic sequence titled "Philosophical Foundations of Medical Education" encouraged reflective discussions about the field's intellectual core.

Since 2022, topics have become explicitly interdisciplinary and future-oriented, such as 3-dimensional atlases, metaverse-based instruction, physician-scientist training, medical artificial intelligence, innovative platforms, and the medical humanities. The 2023–2024 series explored optimizing formative assessment, the 6-year integrated curriculum, student-centered learning, entrepreneurship education, and debriefs from both international and domestic conferences. What began as a narrowly focused pedagogical series has grown into a multilayered forum that includes learners, faculty, and forward-looking strategy. The seminars are held on the first Wednesday of each month at 17:00, providing a regular venue for ongoing professional discussion (Supplement 3).

Leadership camp

Aligned with the College's mission to cultivate "women leaders in health care," a 2-day residential leadership workshop was introduced in 2006 for incoming graduate-entry students. The camp emphasized leadership, communication, and empathy through a combination of orientation activities, campus and historical tours, and interactive sessions. Evening programs paired students with alumnae mentors to transmit institutional heritage and values (Supplement 4).

The final cohort of graduate-entry students completed the program in 2016. Its objectives are now achieved through first-year orientation and the pre-medical course "Team Leadership Project," reflecting the College's ability to adapt to changing educational structures.

Ewha Social Active Communication

ESAC is an annual co-curricular lecture series launched in 2008 to broaden students' professional and communicative perspectives. Speakers from medicine, public health, and diverse social sectors address global, cultural, and policy topics.

Early sessions explored themes such as the World Health Organization's role, traditional medicine, women's health and leadership, and comparative healthcare systems in Vietnam and the

United States. Subsequent years added sessions on the United States Medical Licensing Examination experience, Christian leadership, disability welfare, courageous women, Korean social trends, and international medical aid. Later installments addressed healthcare issues in Azerbaijan, African nations, and other regions.

Since 2014, topics have spanned evolving leadership paradigms, United Nations International Children's Emergency Fund initiatives, pathogen genomics, and German psychiatric practices. Recent sessions—such as US healthcare innovation (2022) and digital health and ecological transitions (2023)—encourage students to rethink physicians' roles in future health systems. ESAC has become a key platform for nurturing globally aware, communication-focused professional identities (Supplement 5).

Problem-based learning

PBL was introduced in 2006 with 2 original modules for first- and second-year medical students. Initially, the cases were based on single diseases, such as hepatitis A or acute appendicitis, but the case library expanded by 2014 to include oncology, immunology, and obstetrics/gynecology topics.

A major update in 2015 introduced complex cases, including premenstrual syndrome, acute cholecystitis, and chronic obstructive pulmonary disease, that required integrative clinical reasoning. Between 2018 and 2019, the curriculum added advanced cardiovascular and respiratory cases (such as stable angina, hemoptysis, cerebral infarction) as well as end-stage colorectal cancer.

However, operational challenges related to securing sufficient tutors, ensuring objective assessment, and managing faculty workload increased over time. As a result, the Curriculum Committee voted on June 16, 2021, to discontinue "PBL I & II." After a period of transition, the last PBL session was delivered during the summer term of 2023.

Team-based learning

Introduced shortly after the Medical Education Office was founded, TBL initially targeted first- and second-year courses. In 2013, a dedicated TBL subcommittee was formed to coordinate new modules, and members visited the Catholic University's MASTER Center to benchmark best practices. Subsequent all-faculty workshops established broad support, and a purpose-built TBL classroom (Medical Building, Room 301) was opened after the move to the Magok campus.

Since 2023, semester-based TBL workshops have become routine, and legacy lectures are being systematically converted into TBL modules. Student feedback is collected and analyzed to drive iterative quality improvement. The current approach focuses on expanding TBL in pedagogically appropriate courses and topics throughout the curriculum, rather than simply increasing the number of TBL sessions (Supplement 6).

Enhancement of student research capacity

A longitudinal Research-Intensive Curriculum integrates research training from the first pre-medical year through the fourth medical year. Foundational courses—such as "Basic Experimental Techniques," "Research Design and Data Analysis," "Manuscript Writing and Presentation," and "Individualized Medical Science Research"—develop core competencies. Advanced electives, including "Advanced Medical Science Research," are available during the early clinical years, while a fourth-year elective allows for full-time laboratory immersion.

A 2024 curriculum reform introduced the spiral "Creative Challenge Research Track," which begins in the first pre-medical semester and continues through the second semester of the second clinical year. Early theoretical training is now offered as "Foundations of Creative Research," and a new course, "Creative Research Challenge," has been introduced (Supplement 7).

Newsletter

Since the inception of the Medical Education Office, a college-wide newsletter has played a vital role in disseminating key educational news and fostering internal communication. Following the establishment of the Ewha Center for Medical Education in 2022, a dedicated Ewha Medical Education Newsletter was launched. Published quarterly in March, June, September, and December, this newsletter covers faculty development, educational innovation, student-centered initiatives, and more. Issues are distributed by email to all campus members, provided in print on campus, and mailed to all 39 Korean medical colleges, thereby enhancing inter-institutional collaboration and visibility (Supplement 8).

Accreditation of medical education

Ewha received full accreditation during the second accreditation cycle in 2009 and was awarded a 5-year "Excellent" rating in 2010. Subsequent comprehensive reviews in 2014 and 2020 resulted in the highest 6-year accreditation status, attesting to the College's continued excellence in educational systems and outcomes. Interim reviews in 2022 and 2024 maintained the 6-year term, underscoring the strength of Ewha's quality assurance processes. These distinctions reflect the College's comprehensive efforts in curriculum development, assessment improvement, faculty development, and infrastructural enhancement, solidifying its national and international reputation.

Conclusion

By systematically analyzing 2 decades of accumulated materials (2004–2025) across chronological phases and thematic domains, this study has illuminated both the developmental trajectory of medical education at Ewha Womans University College of Medicine and the underlying educational philosophy. The archival review revealed significance beyond administrative restructuring—a sustained commitment to embedding Ewha's distinctive identity in every aspect of teaching and learning.

Since the launch of the Medical Education Office in 2004, Ewha's program has advanced through intensive faculty development, curricular innovation, assessment system refinement, and the introduction of integrative, interdisciplinary teaching. At every stage, the dedication and collaboration of faculty members and professional staff have been crucial in shaping and expanding the educational enterprise, resulting in the robust framework in place today. These milestones are not isolated events, but rather represent an organic continuum of institutional growth. What began as the seed of educational practice has—over time—matured into a thriving, interconnected forest. This living forest stands as both a testament to past achievements and a foundation for the future advancement of Ewha's medical education.

Finally, it is important to emphasize that the progress chronicled here reflects not only the accomplishments of the past 20 years, but also the ongoing efforts of countless predecessors who, since the College's founding, have dedicated themselves to cultivating women leaders in Korean healthcare.

ORCID

Ivo Kwon: https://orcid.org/0000-0002-2690-1849
Somi Jeong: https://orcid.org/0000-0002-6984-9753
Seung-Jung Kim: https://orcid.org/0000-0003-4927-2018
Ara Ko: https://orcid.org/0009-0000-1845-8376
Hyeonji Jeon: https://orcid.org/0009-0003-3306-9485

Authors' contribution

Conceptualization: IK, SJK, SMJ. Data curation: ARK. Methodology: SMJ. Formal analysis/validation: SMJ, IK. Project administration: HJJ. Writing–original draft: SMJ, HJJ. Writing–review & editing: SJK, HJJ, ARK.

Conflict of interest

No potential conflict of interest relevant to this article was reported.

Funding

None.

Data availability

Not applicable.

Acknowledgments

None.

Supplementary materials

Supplementary files are available from https://doi.org/10.7910/DVN/REHLFQ

Supplement 1. Commemorative photograph celebrating the 20th anniversary of the Department of Medical Education, Ewha Womans University College of Medicine.

Supplement 2. Principal themes addressed in the Ewha faculty development workshops.

Supplement 3. Major topics covered in the Ewha medical education seminar series.

Supplement 4. Photographs from the Ewha graduate-entry leadership camp.

Supplement 5. List of lecture themes in the Ewha Social Active Communication (ESAC) program.

Supplement 6. Photographic documentation related to team-based learning at Ewha.

Supplement 7. Core team-based learning courses offered at Ewha Womans University College of Medicine.

Supplement 8. Issues of the Ewha Medical Education Newsletter.

References

1. Educational support archives of the Ewha Medical Education Office (2004–2025). Ewha Womans University; 2025.
2. Ewha Womans University College of Medicine self-evaluation reports. Ewha Womans University; 2025.
3. Minutes of the Curriculum Committee, Ewha Womans University College of Medicine (2004–2025). Ewha Womans University; 2025.

eISSN: 2234-2591
Open Access

Original article

Ewha Med J 2025;48(3):e42
https://doi.org/10.12771/emj.2025.00423

The Ewha Medical Journal

Insertion of a G-quadruplex or hairpin structure into the 5’ UTR or poly(A) sequences reduces translation efficiency of the encephalomyocarditis virus internal ribosome entry site: a preclinical *in vitro* study

Yun Ji Kim, So-Hee Hong[*]

Department of Microbiology, Ewha Womans University College of Medicine, Seoul, Korea

Purpose: Internal ribosome entry site (IRES) elements, present in both viral and cellular messenger RNAs (mRNAs), facilitate cap-independent translation by recruiting ribosomes to internal regions of mRNA. This study aimed to investigate the impact of inserting G-quadruplex and hairpin structures into the 5' untranslated region (UTR) and poly(A) sequences on the translation efficiency of the encephalomyocarditis virus (EMCV) IRES, using an IRES-based RNA platform encoding OX40L, 4-1BBL, and GFP.

Methods: G-quadruplex and hairpin structures, derived from HIV-1 (human immunodeficiency virus type 1) or custom-designed, were synthesized and inserted into the 5' UTR and poly(A) tail regions of EMCV IRES vectors. These constructs were amplified by polymerase chain reaction, ligated into plasmids, and transcribed *in vitro*. B16 melanoma, TC-1 tumor, and HEK293 cells were transfected with these RNA constructs. Protein expression levels were assessed at 6, 12, and 24 hours post-transfection by flow cytometry and fluorescence microscopy. Statistical analyses employed one-way analysis of variance with the Dunnett test.

Results: The insertion of G-quadruplex and hairpin structures altered RNA secondary structure, significantly reducing protein expression. In the 5' UTR, the G-quadruplex nearly abolished OX40L expression (1.18%±0.41% at 6 hours vs. 18.23%±0.16% for control), while the hairpin structure reduced it (16.29%±1.46% vs. 22.84%±1.17%). In the poly(A) tail region, both structures decreased GFP expression across all cell lines (4.86%±1.35% to 7.27%±0.32% vs. 39.56%±2.07% in B16 cells).

Conclusion: Inserting G-quadruplex and hairpin structures into EMCV IRES UTRs inhibits translation efficiency, suggesting the need for precise RNA structure modeling to enhance IRES-mediated translation.

Keywords: Encephalomyocarditis virus; Internal ribosome entry sites; Messenger RNA; Transfection

Introduction

Background

Internal ribosome entry sites (IRESs) are RNA elements that recruit ribosomes to internal regions of messenger RNAs (mRNAs), enabling translation through a cap-independent pathway [1,2]. IRESs were first discovered in viruses belonging to the Picornaviridae family, such as poliovirus (PV) and encephalomyocarditis virus (EMCV), in the late 1980s [3]. Subsequently, numerous IRESs have been identified in viral and cellular mRNAs [4]. During viral infection or cellular stress, various mechanisms often suppress conventional cap-dependent translation [1]. Under these adverse conditions, viruses and certain cellular mRNAs rely on IRESs to maintain protein translation [2]. Compared with cap-dependent expression systems, viral IRES platforms offer advantages such as bicistronic gene expression, cap independence, and flexible design suitable for therapeutic applications [5].

The regulation of protein expression is a complex process involving multiple levels of control [6]. Recent studies have empha-

[*]Corresponding email: shhong13@ewha.ac.kr

Received: April 25, 2025 Accepted: May 7, 2025

sized the crucial role of secondary RNA structural elements in fine-tuning translation efficiency [7,8]. Among these elements, G-quadruplexes and hairpin structures have gained particular attention due to their ability to improve protein production under various conditions, such as infection, hypoxia, and DNA damage [2,3].

G-quadruplexes are 4-stranded helical structures formed in guanine-rich, single-stranded RNA (ssRNA) sequences. They function similarly to IRESs and promote cap-dependent translation initiation [9-11]. The G-quadruplex follows a structural algorithm characterized by the sequence GXN1-7GXN1-7GXN1-7GXN1-7 (X ≥ 3, N represents any nucleotide), consisting of guanine quartets (G-quartets) surrounded by loop regions [12]. G-quadruplex structures are frequently found in proto-oncogenes related to cancer, such as c-myc, VEGF, and Bcl-2 [13]. These structures are thermodynamically stable, easily formed, and contribute to RNA stability [12,14]. Under cellular stress conditions, G-quadruplex structures help maintain RNA-ribosome interactions, thereby supporting continuous protein translation [10]. They are also well known for influencing gene splicing and enhancing the binding of specific transcription factors [10,13]. A recent study has shown that increasing the loop size of G-quadruplexes enhances protein translation [6].

The RNA hairpin structure resembles a loop or U-shape and forms when 2 regions of the same RNA strand base-pair to create a double helix [15]. Under specific conditions, such as hypoxia and DNA damage, hairpin structures can enhance start codon recognition, thereby increasing protein expression [7,16]. Hairpin structures also contribute to RNA stability by forming stable secondary structures at mRNA termini, preventing degradation, and extending RNA half-life, which ultimately enhances overall protein expression [17,18]. Notably, the use of multiple hairpin structures rather than a single structure yields superior translational outcomes [17].

Additionally, other studies have demonstrated that the combined use of hairpins and G-quadruplex structures significantly enhances translation efficiency [6,17].

Objectives

Based on this evidence, in the present study, we designed and inserted G-quadruplexes and hairpin structures into the EMCV IRES and assessed changes in the protein expression levels of the IRES-encoded gene to reveal the effects of these inserted structures on the translational function of the IRES.

Methods

Ethics statement

This study constitutes laboratory research, not involving human subjects. Therefore, neither institutional review board approval nor informed consent was required.

Design and synthesis of G-quadruplex and hairpin structures

Plasmids encoding 4-1BBL, OX40L, and GFP, based on the EMCV IRES, were used as the backbone, and these vectors were constructed according to previously described cloning methods [5]. The G-quadruplex and hairpin sequences were designed based on human immunodeficiency virus type 1 (HIV-1)-derived sequences, with fundamental structural details and custom-designed sequences described in Table 1.

All genes for structural design were synthesized by Cosmo Genetech Inc. (https://www.cosmogenetech.com). The inserted G-quadruplex and hairpin genes were amplified by polymerase chain reaction (PCR) using the following reaction conditions: 10 ng of plasmid template, 0.5 μL of 10 pmol forward primer, 0.5 μL of 10 pmol reverse primer, 4 μL of Phusion 5× buffer, 0.4 μL of dNTPs, and 0.2 μL of Phusion enzyme (Thermo Fisher Scientific; https://www.thermofisher.com).

The following primers were used for PCR:
E-H forward: 5′- CCGAATTCTAATACGACTCACTAT -3′
E-H reverse: 5′- CCCCTAGGAATGCTCGTCAAG -3′
E-G forward: 5’- CCGAATTCTAATACGACTCACTAT -3’
E-G reverse: 5’- CCCCTAGGAATGCTCGTCAAG -3’
E-M1 forward: 5′- CCGTCGACCGATCGTAGTGTAGTCAC -3′
E-M1 reverse: 5′- CCGCGGCCGCGCTAGC -3′
E-M2 forward: 5′- CCGTCGACCGATCGTAGTGTAGTCAC -3′
E-M2 reverse: 5′- CCGCGGCCGCTGGTAATG -3′
E-M3 forward: 5′- CCGTCGACCGATCGTAGTGTAGTCAC -3′
E-M3 reverse: 5′- CCGCGGCCGCCAGGCT -3′
E-M4 forward: 5’- CCGTCGACCGATCGTAGTGTAGTCAC -3’
E-M4 reverse: 5’- CCGCGGCCGCCAGGCT -3’

The PCR conditions were as follows: initial denaturation at 98°C for 30 seconds, followed by 35 cycles of denaturation at 98°C for 10 seconds, annealing at 55°C for 10 seconds, and extension at 72°C for 20 seconds. A final extension was performed at

Table 1. Sequence of structural elements

Abbreviation	Structure	Origin	Sequence (+linker)	Total length (including T7–Not I) (nt)	Position of the structural element
5' UTR modification					
E-G	G-quadruplex	HIV-1	CCAGGGAGGCGTGGCCTGGGCG GGACTGGGGAGTGGCGAG	1488	25–52
E-H	Hairpin	HIV-1	GGTCTCTCTGGTTAGA CCAGAGAGCC	1824	25–55
Poly(A) tail region modification					
E-M1	G-quadruplex	Designed	GGGAAAGGGUUUGGGAAAGGG	1638	1594–1630
E-M2	G-quadruplex + hairpin	Designed	GGGAAAGGGUUUGGGAAAGG GGAAGATCAAG**GCTAGC**ACCA UUACCCGCCUUGGGUAAUGG UUGGUAAUGGGUUCCGCCCAU UACCA	1688	1594–1680
E-M3	G-quadruplex + hairpin	HIV-1	CCAGGGAGGCGTGGCCTGGG CGGGACTGGGGAGTGGCGAG **GCTAGC**GGTCTCTCTGGTTAG ACCAGATCTGAGCCTG	1678	1594–1670
E-M4	G-quadruplex + hairpin	Designed + HIV-1	GGGAAAGGGUUUGGGAAAG GG**GCTAGC**ACCAUUACCCGCC UUGGGUAAUGGUGGTCTCTC TGGTTAGACCA GATCTGAGCCTG	1684	1594–1676

Bold text within the "Sequence (+linker)" column indicates the linker region.
nt, nucleotide; UTR, untranslated region; HIV-1, human immunodeficiency virus type 1.

72°C for 1 minute, followed by indefinite storage at 12°C. Amplified products were analyzed using 2% agarose gel electrophoresis, and product sizes were verified using a 50 bp DNA ladder (Dyne LoadingSTAR+50 bp DNA Ladder; Dynebio; http://www.dynebio.co.kr). Target bands were excised and purified using a gel cleanup kit. The purified gene fragments were digested with respective restriction enzymes (Enzynomics; https://www.enzynomics.com) at 37°C for 1 hour and 30 minutes, followed by enzyme inactivation at 65°C for 30 minutes. After additional purification, the amplified genes were ligated into the pALpA_EMCV IRES vector using T4 ligase (RBC Rapid Ligation Kit, RBC, Taiwan) and incubated at 4°C overnight. The ligation products were transformed into Escherichia coli DH5α (Enzynomics) by heat shock at 42°C. Ampicillin-resistant colonies were selected and cultured in LB medium (Duchefa; https://www.duchefa-farma.com) containing ampicillin (Duchefa). Finally, plasmid DNA was purified using a Plasmid DNA Miniprep S&V kit (Bionics; https://www.bionicsro.co.kr). The final plasmids were confirmed by electrophoresis following digestion with restriction enzymes.

In vitro transcription

DNA templates were linearized using the NotI restriction enzyme (Enzynomics). *In vitro* transcription was performed using the EZ(TM) T7 High Yield In-Vitro Transcription Kit (Enzynomics), driven by the T7 promoter. A 1 μg linearized DNA template was mixed with T7 transcription buffer, $MgCl_2$, 10 mM dithiothreitol, enhancer solution, 5 mM ribonucleoside triphosphates, 200 U of T7 polymerase mix, and ultrapure water, reaching a final reaction volume of 20–100 μL. The reaction mixture was incubated at 37°C for 4–6 hours.

Following transcription, DNA was removed by DNase I (Promega; https://promega.com) treatment at 37°C for 30 minutes. RNA was precipitated using lithium chloride, and double-stranded RNA was eliminated via cellulose purification following previously established methods [19]. RNA purity and concentration were assessed using a NEO-Nabi UV-VIS Nano spectrophotometer (MicroDigital Co. Ltd.; https://www.md-best.com). Only samples with 260/230 and 260/280 absorbance ratios > 1.9 were used for subsequent analyses. RNA was mixed with denaturing dye, heated at 70°C for 10 minutes, and analyzed by electrophoresis on a 1.5% agarose gel. RNA was stained with RedSafe Nucleic Acid Staining Solution, and quality was assessed using the RiboRuler High Range RNA Ladder (Thermo Fisher Scientific).

Cell culture

Mouse B16 melanoma (CRL-6475; ATCC; https://www.atcc.org) and HEK293 cells (Korean Cell Line Bank; https://cellbank.snu.ac.kr) were cultured in Dulbecco's modified Eagle's medium (DMEM; GenDEPOT; https://gendepot.com) supplemented with 10% fetal bovine serum (FBS; Welgene; https://www.welgene.com) and 1% penicillin/streptomycin (Welgene).

Mouse TC-1 tumor cells (CRL-2493; ATCC) were maintained in RPMI 1640 medium (Welgene) supplemented with 10% FBS and 1% penicillin/streptomycin. All cells were incubated at 37°C in a humidified atmosphere containing 5% CO_2.

Transfection

B16, TC-1, and HEK293 cells (7×10^5 cells/well) were seeded into 6-well plates (SPL; http://www.spllifesciences.com) and cultured in DMEM or RPMI 1640 medium (supplemented with 10% FBS and 1% penicillin/streptomycin) at 37°C with 5% CO_2 for 12 hours. After incubation, cells were washed twice with cold phosphate-buffered saline (PBS) and transfected with 5 μg of RNA using Lipofectamine 3000 (Thermo Fisher Scientific) in Opti-MEM (Gibco, Thermo Fisher Scientific) and serum-free medium. Protein expression was assessed by flow cytometry and live imaging using a Leica Thunder Imager (Leica; https://www.leica-microsystems.com).

Flow cytometry analysis

Cells were harvested and resuspended in flow cytometry buffer (PBS containing 1% BSA and 0.01% NaN_3). Fc receptors were blocked by incubation with anti-mouse CD16/32 (TruStain FcX, BioLegend) at 4°C for 15 minutes. Subsequently, cells were stained at 4°C for 30 minutes in the dark with antibodies and dye: anti-mouse CD275 (ICOS Ligand, clone HK5.3, BioLegend), anti-4-1BBL (CD137L, clone TKS-1, BioLegend), anti-CD252 (OX40L, clone RM143L, BioLegend), and LIVE/DEAD Fixable Aqua Dead Cell Stain (Invitrogen).

Statistical methods

Statistical significance was evaluated using one-way analysis of variance followed by the Dunnett post hoc multiple comparisons test. A P-value less than 0.05 was considered statistically significant ($P < 0.05$, $P < 0.01$, $P < 0.001$). Data are expressed as mean ± standard deviation (SD). Analyses were conducted using GraphPad Prism ver. 10.0 (GraphPad Software).

Results

Design and insertion of G-quadruplex and hairpin structures in IRES-based vectors: impact on IRES RNA secondary structure

In this study, 2 different G-quadruplex and hairpin structures were employed: one derived from an HIV-1 sequence and another custom-designed specifically for this study (Fig. 1A, B). To investigate the effects of inserting these structures into the untranslated regions (UTRs) of an IRES-based platform, they were integrated into an EMCV IRES-based RNA vector (Fig. 1C). First, RNA secondary structures derived from HIV-1, previously described in earlier studies [13,20], were synthesized and inserted into the 5' UTR of the EMCV IRES, encoding the co-stimulatory molecules OX40L and 4-1BBL. These secondary structures were positioned between the T7 promoter and the IRES sequence (Fig. 1D, E). The structural elements were directly linked to the T7 promoter and IRES without a spacer or linker. Structural prediction analyses indicated that inserting the hairpin motif minimally impacted the native IRES structure, whereas inserting the G-quadruplex significantly altered the RNA secondary structure (Fig. 1C–E).

To explore the impact of structures located in the 3' UTR on protein expression, G-quadruplex and hairpin motifs were also inserted into the 3' UTR of a GFP-encoding EMCV IRES platform (Fig. 1F–I). The secondary structures were placed downstream of the poly(A) tail, consisting of 100 adenine residues, followed by an additional 10-nucleotide poly(A) sequence. A 12-nucleotide linker sequence was also incorporated to ensure proper spatial positioning of the secondary structures. Hairpin and G-quadruplex structures with distinct sequences were designed and inserted into the same location within the 3' UTR (Fig. 1F–I). Despite sequence differences between the 2 hairpin and 2 G-quadruplex structures, structural prediction analyses demonstrated that all inserted RNA secondary constructs significantly altered the RNA conformation compared to the original vector (Fig. 1C, 1F–I).

Reduced protein expression in EMCV IRES-based RNA platforms facilitated by G-quadruplex and hairpin insertion

To evaluate the impact of G-quadruplex and hairpin structures on translation efficiency, EMCV-IRES platforms containing these secondary structures inserted into either the 5' UTR or poly(A) tail region were transfected into various cell lines. The expression levels of encoded genes such as OX40L, 4-1BBL, or GFP were then evaluated. First, to assess the effect of structures located in the 5' UTR, EMCV-IRES encoding OX40L and 4-1BBL with in-

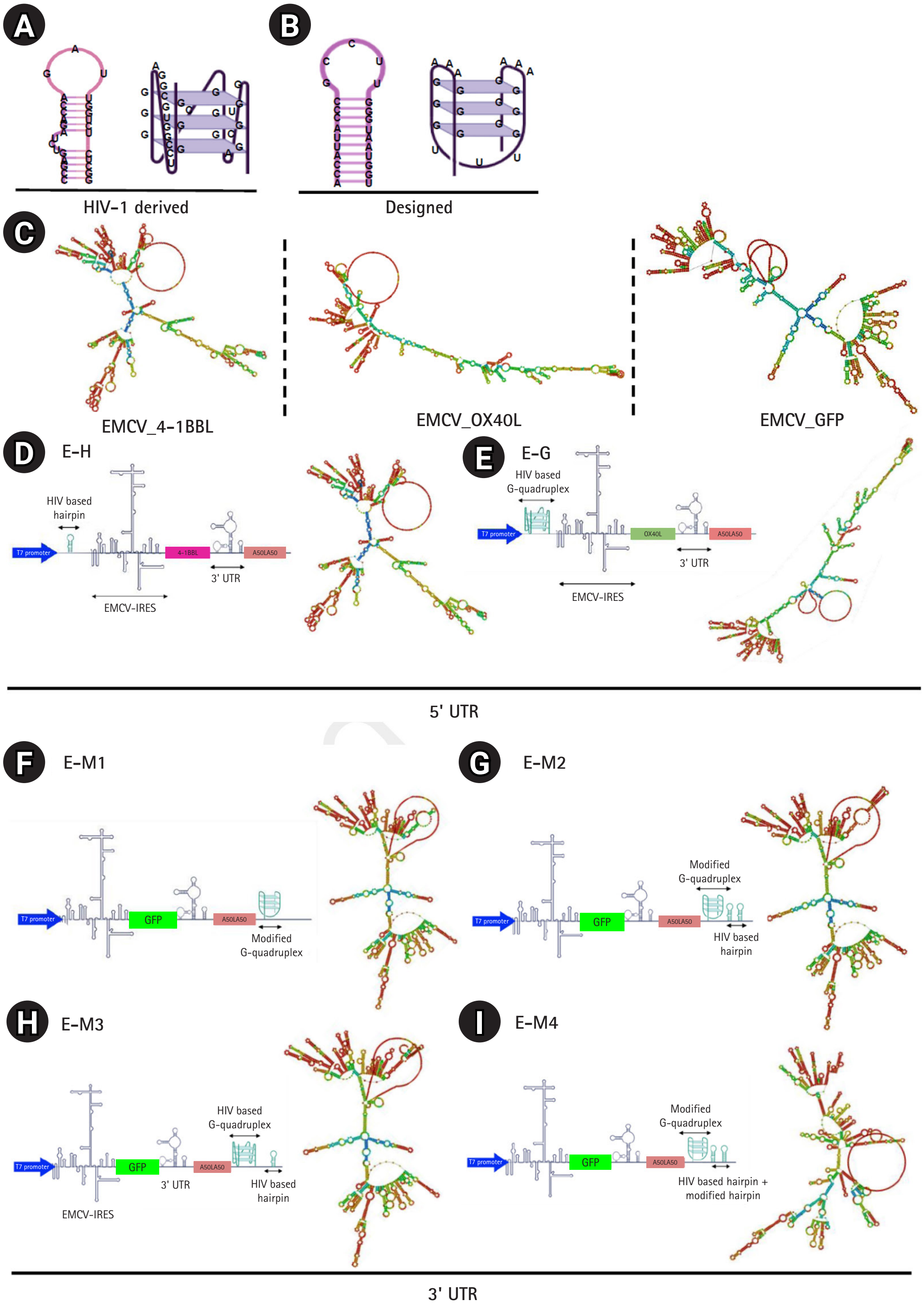


Fig. 1. Design and positioning of secondary RNA structures inserted into the encephalomyocarditis virus (EMCV) internal ribosome entry site (IRES) and their resulting secondary structures. (A) Structure of human immunodeficiency virus type 1 (HIV-1) derived hairpin and G-quadruplex. (B) Structure of designed hairpin and G-quadruplex. (C) Predicted structure of the EMCV IRES-based RNA platform encoding 4-1BBL, OX40L, or GFP using RNAfold program. (D) Design and predicted 2-dimensional (2D) structure of the EMCV IRES-based RNA platform expressing 4-1BBL and containing a hairpin structure in the 5′ untranslated region (UTR). (E) Design and predicted 2D structure of the EMCV IRES-based RNA platform expressing OX40L and containing a G-quadruplex structure in the 5′ UTR. (F–I) Design and predicted 2D structure of the EMCV IRES-based RNA platform expressing GFP and containing a secondary structure in the poly(A) tail.

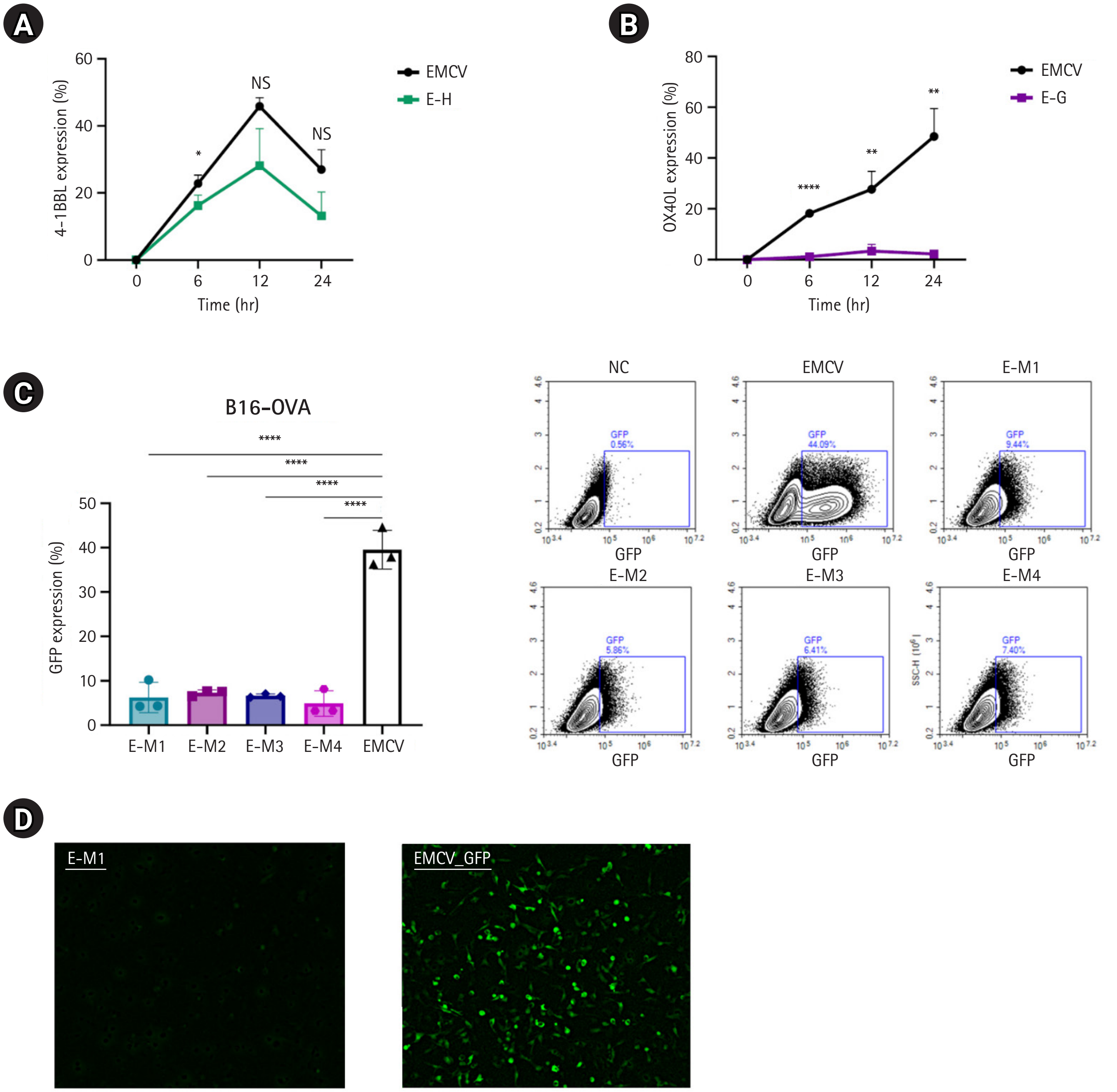


Fig. 2. Reduced protein expression efficiency of encephalomyocarditis virus (EMCV)-internal ribosome entry site (IRES) platforms containing hairpin or G-quadruplex secondary structures. (A) Time-dependent expression levels of 4-1BBL in B16 melanoma cells transfected with EMCV-IRES-4-1BB with or without a hairpin in the 5′ untranslated region (UTR). (B) Time-dependent expression levels of OX40L in B16 melanoma cells transfected with EMCV-IRES-OX40L with or without a G-quadruplex structure in the 5′ UTR. (C) Results of flow cytometry 24 hours after transfection of the B16 melanoma cell line with EMCV-IRES-GFP, with or without structural elements in the poly(A) tail. (D) Live-cell fluorescence imaging of B16 melanoma cells 24 hours post-transfection with EMCV-IRES-GFP constructs, with (E-M1) or without structural elements in the poly(A) tail (EMCV-GFP). (E) Results of flow cytometry 24 hours after transfection of the TC-1 cell line with EMCV-IRES-GFP, with or without structural elements in the poly(A) tail. (F) Results of flow cytometry 24 hours after transfection of the HEK 293 cell line with EMCV-IRES-GFP, with or without structural elements in the poly(A) tail. NS, not significant. (Continued on the next page.)

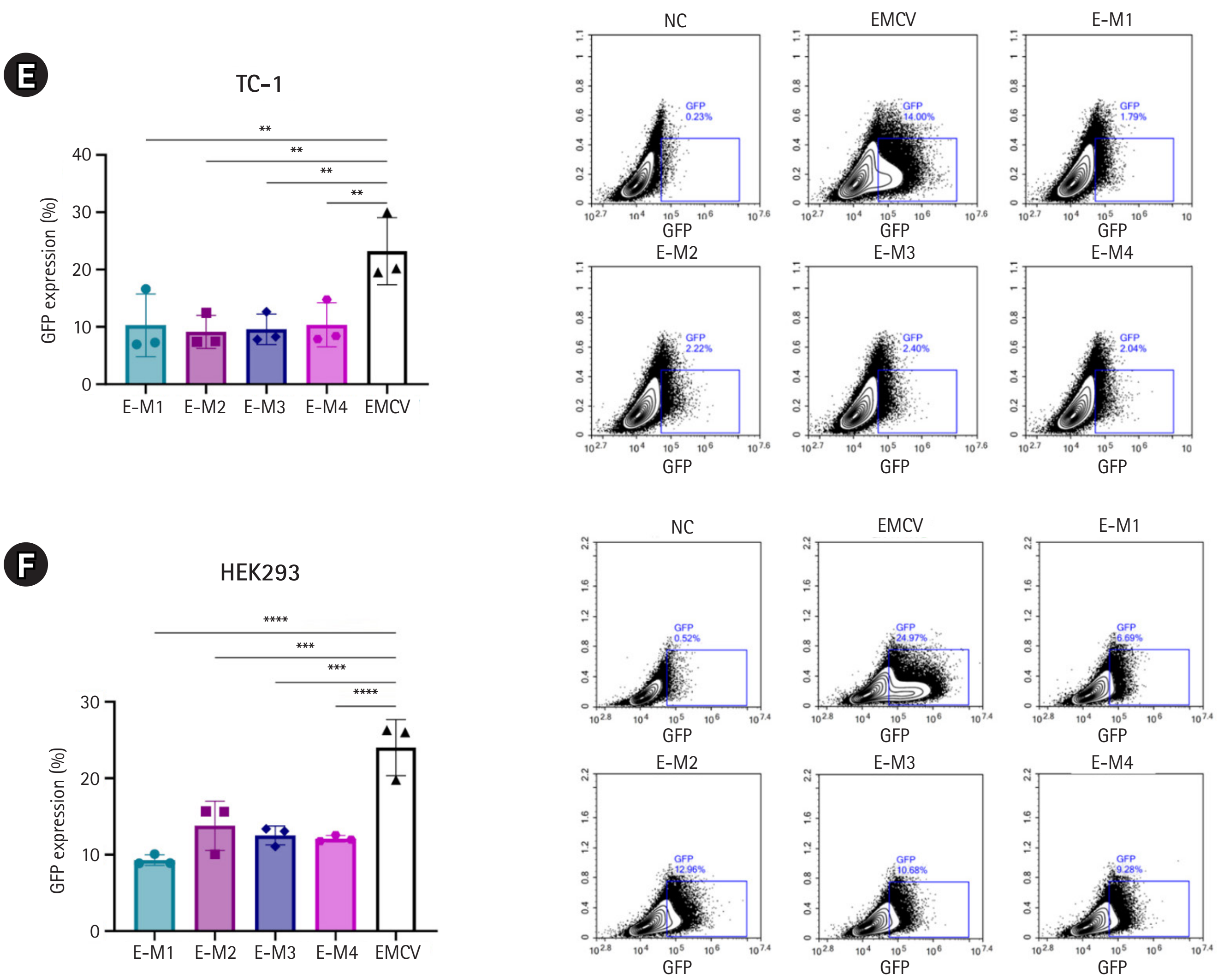


Fig. 2. (Continued; caption shown on previous page).

serted hairpin or G-quadruplex motifs at the 5' UTR terminal were synthesized. These constructs were transfected into B16 melanoma cells, and protein expression levels were measured at 6, 12, and 24 hours post-transfection. Compared to cells transfected with the EMCV-IRES construct lacking these secondary structures, significantly reduced protein expression levels were observed (Fig. 2A, B). Specifically, EMCV-IRES constructs containing hairpin structures resulted in decreased protein expression. At 6 hours post-transfection, EMCV control exhibited a protein expression rate (mean ± SD) of 22.84% ± 1.17%, whereas the E-H construct displayed a reduced rate of 16.29% ± 1.46%. At 12 hours, the expression levels were 45.8% ± 1.21% (EMCV) and 28.13% ± 5.2% (E-H), and at 24 hours, they were 26.98% ± 2.77% and 13.17% ± 3.34%, respectively (Fig. 2A). Notably, insertion of the HIV-1 G-quadruplex structure at the 5' UTR terminal nearly abolished OX40L expression (Fig. 2B). At 6 hours post-transfection, EMCV exhibited a protein expression rate of 18.23% ± 0.16%, compared with only 1.18% ± 0.41% for the E-G construct. At 12 hours, the expression levels were 27.67% ± 3.33% and 3.34% ± 1.28%, and at 24 hours, they were 48.5% ± 5.19% and 2.23% ± 0.47%, respectively (Fig. 2B).

To determine whether these secondary structures affect protein expression regardless of their location within the UTRs, hairpin and G-quadruplex motifs were inserted downstream of the poly(A) tail in the 3' UTR of an EMCV IRES-based vector encoding GFP. These EMCV-IRES constructs were transfected into B16 melanoma, TC-1 tumor, and HEK293 human embryonic kidney cells, and GFP expression was analyzed at 24 hours post-transfection. Consistent with results observed in the 5' UTR experiments, all ssRNA constructs containing secondary structures at the poly(A) tail terminal of the 3' UTR exhibited significantly reduced GFP expression compared to the control EMCV

IRES (Fig. 2C–F). In B16 cells, the EMCV IRES control exhibited a protein expression rate (mean ± SD) of 39.56% ± 2.07%, whereas E-M1, E-M2, E-M3, and E-M4 showed notably lower expression levels: 6.23% ± 1.61%, 7.27% ± 0.32%, 6.6% ± 0.19%, and 4.86% ± 1.35%, respectively (Fig. 2C). Fluorescence microscopy confirmed these findings, showing clearly visible GFP fluorescence only in the EMCV IRES control group, whereas the E-M constructs exhibited minimal fluorescence, indicating extremely low protein expression (Fig. 2D). Similar results were obtained across other cell lines. In TC-1 cells, the EMCV IRES control showed an expression rate of 23.22% ± 2.76%, whereas the E-M1, E-M2, E-M3, and E-M4 constructs exhibited significantly lower expression levels: 10.28% ± 2.59%, 9.13% ± 1.37%, 9.57% ± 1.26%, and 10.35% ± 1.82%, respectively (Fig. 2E). Likewise, in HEK293 cells, the EMCV control exhibited protein expression of 24% ± 1.73%, compared with significantly reduced levels in E-M1, E-M2, E-M3, and E-M4 constructs: 9.27% ± 0.33%, 13.76% ± 1.52%, 12.5% ± 0.59%, and 12.09% ± 0.2%, respectively (Fig. 2F). These findings consistently demonstrated that G-quadruplex and hairpin insertions significantly reduce protein expression across different cell types (Fig. 2C–F).

Discussion

Key results

This study demonstrated that inserting G-quadruplex and hairpin structures into the 5' UTR and poly(A) region of an EMCV IRES-based RNA platform significantly alters RNA secondary structure and reduces protein expression. In the 5' UTR, the insertion of the G-quadruplex nearly abolished OX40L expression, whereas insertion of the hairpin structure reduced expression to a lesser extent. In the 3' UTR, both types of structures led to decreased GFP expression across B16, TC-1, and HEK293 cell lines, with expression levels dropping to as low as 4.86% compared to control constructs. Based on these results, the study suggests that insertion of G-quadruplex and hairpin motifs reduces translational efficiency in EMCV IRES-based vectors, irrespective of whether they are positioned within the 5' or 3' UTR.

Interpretation/comparison with previous studies

IRES sequences facilitate cap-independent translation, enabling the expression of multiple genes from a single mRNA transcript [5]. This characteristic renders IRES elements valuable tools for RNA-based therapeutics targeting diverse diseases. In this study, we inserted RNA secondary structures to potentially increase the translation efficacy of the EMCV IRES. Previous reports indicated that RNA secondary structures, such as G-quadruplexes and hairpins, enhance RNA stability and translation efficiency [6,21]. In contrast, we observed that the incorporation of these secondary structures into the EMCV IRES-based vector resulted in decreased translation efficiency. This reduction in protein expression consistently occurred regardless of whether the structures were positioned at the 5' terminus or downstream of the poly(A) tail.

One possible explanation for the discrepancy between our findings and previous reports is that the hairpin structure utilized in our study may have been too small to effectively protect the mRNA terminus. Additionally, the absence of a spacer sequence between the inserted hairpin and the IRES element could have disrupted proper IRES functionality, thereby diminishing translational efficiency.

Since numerous translation-associated factors bind to secondary structures within the IRES, the insertion of hairpins or G-quadruplexes might interfere with factor binding or induce conformational changes in the 3-dimensional architecture of the IRES-based platform. This could potentially hinder ribosomal accessibility or compromise the structural integrity of the IRES, resulting in reduced translation.

Limitations/suggestions for further studies

In this study, our analysis focused exclusively on the EMCV IRES-based platform because of its previously established excellent translational capabilities. Therefore, further investigations using alternative viral and cellular IRES elements are needed to determine whether inserting these RNA secondary structures similarly reduces translation efficiency across other IRES platforms. Generally, IRES-based platforms exhibit lower translation efficiency compared to cap-dependent systems. Therefore, strategies such as stabilizing the IRES, improving the recruitment of translation-associated factors, or enhancing interactions between the IRES and these factors could enhance IRES translation efficiency. Although RNA structure insertion reduced translation in our current platform, the potential of RNA secondary structures to improve translational efficiency still exists. More precise and strategic design and positioning of RNA structures are required to optimize IRES-based platforms.

Implications

Enhancement of translational efficiency through RNA structural modifications could facilitate the development of efficient mRNA-based therapeutics, potentially enabling reduced mRNA dosages. This approach may minimize potential side effects and improve cost-effectiveness in therapeutic applications.

Conclusion

Inserting hairpin or G-quadruplex structures upstream of the 5' UTR or downstream of the poly(A) tail significantly reduced the translation efficiency of EMCV IRES-encoded genes. To effectively enhance translation in IRES platforms, precise 2-dimensional or 3-dimensional structural modeling is required to ensure that inserted RNA structures do not disrupt the native IRES conformation.

ORCID

Yun Ji Kim: https://orcid.org/0009-0001-6842-4446
So-Hee Hong: https://orcid.org/0000-0002-2833-8025

Authors' contributions

Conceptualization: SHH. Methodology/formal analysis/validation: YJK. Project administration: SHH. Funding acquisition: SHH. Writing–original draft: YJK. Writing–review & editing: YJK, SHH.

Conflict of interest

No potential conflict of interest relevant to this article was reported.

Funding

This study was supported by a grant from the Ministry of Science and ICT (MIST; grant number 2022M3E5F1016595).

Data availability

The data supporting the findings of this study are available from the corresponding author upon request.

Acknowledgments

None.

Supplementary materials

None.

References

1. Yang Y, Wang Z. IRES-mediated cap-independent translation, a path leading to hidden proteome. J Mol Cell Biol 2019;11:911-919. https://doi.org/10.1093/jmcb/mjz091
2. Spriggs KA, Bushell M, Mitchell SA, Willis AE. Internal ribosome entry segment-mediated translation during apoptosis: the role of IRES-trans-acting factors. Cell Death Differ 2005;12:585-591. https://doi.org/10.1038/sj.cdd.4401642
3. Jang SK, Krausslich HG, Nicklin MJ, Duke GM, Palmenberg AC, Wimmer E. A segment of the 5' nontranslated region of encephalomyocarditis virus RNA directs internal entry of ribosomes during in vitro translation. J Virol 1988;62:2636-2643. https://doi.org/10.1128/JVI.62.8.2636-2643.1988
4. Li Y, Zhang L, Wang L, Li J, Zhao Y, Liu F, Wang Q. Structure and function of type IV IRES in picornaviruses: a systematic review. Front Microbiol 2024;15:1415698. https://doi.org/10.3389/fmicb.2024.1415698
5. Ko HL, Park HJ, Kim J, Kim H, Youn H, Nam JH. Development of an RNA expression platform controlled by viral internal ribosome entry sites. J Microbiol Biotechnol 2019;29:127-140. https://doi.org/10.4014/jmb.1811.11019
6. Lee CY, Joshi M, Wang A, Myong S. 5'UTR G-quadruplex structure enhances translation in size dependent manner. Nat Commun 2024;15:3963. https://doi.org/10.1038/s41467-024-48247-8
7. Leppek K, Das R, Barna M. Functional 5' UTR mRNA structures in eukaryotic translation regulation and how to find them. Nat Rev Mol Cell Biol 2018;19:158-174. https://doi.org/10.1038/nrm.2017.103
8. Chiaruttini C, Guillier M. On the role of mRNA secondary structure in bacterial translation. Wiley Interdiscip Rev RNA 2020;11:e1579. https://doi.org/10.1002/wrna.1579
9. Spiegel J, Adhikari S, Balasubramanian S. The structure and function of DNA G-quadruplexes. Trends Chem 2020;2:123-136. https://doi.org/10.1016/j.trechm.2019.07.002
10. Shu H, Zhang R, Xiao K, Yang J, Sun X. G-quadruplex-binding proteins: promising targets for drug design. Biomolecules 2022;12:648. https://doi.org/10.3390/biom12050648
11. Hansel-Hertsch R, Spiegel J, Marsico G, Tannahill D, Balasubramanian S. Genome-wide mapping of endogenous G-quadruplex DNA structures by chromatin immunoprecipitation and high-throughput sequencing. Nat Protoc 2018;13:551-564. https://doi.org/10.1038/nprot.2017.150
12. Lee DS, Ghanem LR, Barash Y. Integrative analysis reveals RNA G-quadruplexes in UTRs are selectively constrained and enriched for functional associations. Nat Commun 2020;11:527. https://doi.org/10.1038/s41467-020-14404-y
13. Sissi C, Gatto B, Palumbo M. The evolving world of protein-G-quadruplex recognition: a medicinal chemist's perspective. Biochimie 2011;93:1219-1230. https://doi.org/10.1016/j.biochi.2011.04.018
14. Fracchioni G, Vailati S, Grazioli M, Pirota V. Structural unfolding of G-quadruplexes: from small molecules to antisense strategies. Molecules 2024;29:3488. https://doi.org/10.3390/molecules29153488
15. Kiliszek A, Blaszczyk L, Kierzek R, Rypniewski W. Stabilization

of RNA hairpins using non-nucleotide linkers and circularization. Nucleic Acids Res 2017;45:e92. https://doi.org/10.1093/nar/gkx122

16. Bao C, Zhu M, Nykonchuk I, Wakabayashi H, Mathews DH, Ermolenko DN. Specific length and structure rather than high thermodynamic stability enable regulatory mRNA stem-loops to pause translation. Nat Commun 2022;13:988. https://doi.org/10.1038/s41467-022-28600-5
17. Solodushko V, Fouty B. Terminal hairpins improve protein expression in IRES-initiated mRNA in the absence of a cap and polyadenylated tail. Gene Ther 2023;30:620-627. https://doi.org/10.1038/s41434-023-00391-4
18. Heinicke LA, Wong CJ, Lary J, Nallagatla SR, Diegelman-Parente A, Zheng X, Cole JL, Bevilacqua PC. RNA dimerization promotes PKR dimerization and activation. J Mol Biol 2009;390:319-338. https://doi.org/10.1016/j.jmb.2009.05.005
19. Baiersdorfer M, Boros G, Muramatsu H, Mahiny A, Vlatkovic I, Sahin U, Kariko K. A facile method for the removal of dsRNA contaminant from in vitro-transcribed mRNA. Mol Ther Nucleic Acids 2019;15:26-35. https://doi.org/10.1016/j.omtn.2019.02.018
20. Piekna-Przybylska D, Sullivan MA, Sharma G, Bambara RA. U3 region in the HIV-1 genome adopts a G-quadruplex structure in its RNA and DNA sequence. Biochemistry 2014;53:2581-2593. https://doi.org/10.1021/bi4016692
21. Solodushko V, Kim JH, Fouty B. A capless hairpin-protected mRNA vaccine encoding the full-length Influenza A hemagglutinin protects mice against a lethal Influenza A infection. Gene Ther 2025 Feb 23 [Epub]. https://doi.org/10.1038/s41434-025-00521-0

eISSN: 2234-2591
Open Access

Original article

Ewha Med J 2025;48(3):e43
https://doi.org/10.12771/emj.2025.00640

The Ewha Medical Journal

Spatiotemporal associations between air pollution and emergency room visits for cardiovascular and cerebrovascular diseases in Korea using a multivariate graph autoencoder modeling approach: an ecological study

Sohee Wang[1,2], Seungpil Jeong[3*], Eunhee Ha[1,2,4*]

[1] *Department of Environmental Medicine, Ewha Womans University College of Medicine, Seoul, Korea*
[2] *Graduate Program in System Health Science and Engineering, Ewha Womans University College of Medicine, Seoul, Korea*
[3] *Convergence Medical Research Institute, Ewha Womans University Mokdong Hospital, Seoul, Korea*
[4] *Institute of Ewha-SCL for Environmental Health (IESEH), Ewha Womans University College of Medicine, Seoul, Korea*

Purpose: This study aimed to assess the spatiotemporal associations between air pollution and emergency room visits for cardiovascular and cerebrovascular diseases in South Korea using a graph autoencoder (GAE). A multivariate graph-based approach was used to uncover seasonal and regional variations in pollutant–disease relationships.

Methods: We collected monthly data from 2022 to 2023, including concentrations of 6 air pollutants (SO_2, NO_2, O_3, CO, PM_{10}, and $PM_{2.5}$) and emergency room visits for 4 disease types: cardiac arrest, myocardial infarction, ischemic stroke, and hemorrhagic stroke. Pearson correlation coefficients were used to construct adjacency matrices, which, along with normalized feature matrices, were used as inputs to the GAE. The model was trained separately for each month and region to estimate the strength of pollutant–disease associations.

Results: The pollutant–disease network structures exhibited clear seasonal variations. In winter, strong associations were observed between O_3, NO_2, and all disease outcomes. In spring, $PM_{2.5}$ and PM_{10} were strongly linked to cardiac and stroke-related visits. These connections weakened during summer but became more pronounced in autumn, especially for NO_2 and cardiac arrest. Urban areas displayed denser and stronger associations than non-urban areas.

Conclusion: Our findings underscore the necessity for season- and region-specific air quality management strategies. In winter, focused control of O_3 and NO_2 is needed in urban areas, while in spring, PM mitigation is required in urban and selected rural regions. Autumn NO_2 control may be especially beneficial in non-urban areas. Spatiotemporally tailored interventions could reduce the burden of air pollution-related emergency room visits.

Keywords: Air pollution; Cardiovascular diseases; Environmental pollutants; Epidemiology; Ischemic stroke; Republic of Korea

Introduction

Background

Cardiovascular and cerebrovascular diseases are among the leading causes of morbidity and mortality worldwide. In South Korea, their burden is steadily increasing due to rapid urbanization, population aging, and lifestyle changes [1]. Traditionally, risk factors such as hypertension, diabetes, dyslipidemia, smoking, and physical inactivity have been the main focus of disease prevention and management efforts. However, in recent years, environmental influences, particularly air pollution, have attracted increasing attention as significant and modifiable contributors to disease onset and progression [2].

Numerous epidemiological studies have reported that both

*Corresponding email: 01637@eumc.ac.kr (SJ), eunheeha@ewha.ac.kr (EH)
Received: July 3, 2025 Revised: July 16, 2025 Accepted: July 17, 2025

*These authors contributed equally to this work as corresponding authors.

short- and long-term exposure to ambient pollutants—including fine particulate matter ($PM_{2.5}$ and PM_{10}), nitrogen dioxide (NO_2), and sulfur dioxide (SO_2)—is associated with elevated risks of myocardial infarction (MI), ischemic stroke (IS), and cardiac arrest (CA) [2-5]. For instance, time-series analyses and cohort studies conducted in Korea and globally have demonstrated significant associations between pollutant concentration spikes and increased hospital admissions or mortality due to cardiovascular events [6-8]. Nevertheless, these studies often employ traditional statistical methods, which primarily focus on single-variable relationships or use linear models to estimate risk ratios or odds ratios [6-8]. While such approaches provide valuable insights into individual pollutant effects, they have limitations in addressing the complex, nonlinear, and multivariate nature of real-world environmental exposures and disease dynamics [9].

Moreover, both air pollution and health outcomes show pronounced spatiotemporal variation. Pollutant concentrations fluctuate seasonally due to meteorological and atmospheric factors, and regionally owing to differences in industrial activity, traffic density, and urban infrastructure. At the same time, the health impacts of pollution may vary depending on regional healthcare access, population vulnerability, and baseline disease prevalence [10-12]. These multifaceted interactions prompt important questions: How do pollutant–disease relationships vary across seasons and geographic regions? Which pollutants are most strongly associated with cardiovascular or cerebrovascular diseases in specific contexts?

To address these questions, more advanced analytical tools are needed. Recent developments in artificial intelligence and deep learning, especially within the field of graph neural networks, offer new possibilities for modeling complex relationships between interconnected variables. Among these, the graph autoencoder (GAE) has shown promise in learning hidden structures and association strengths from multivariate graph data [13].

Objectives

In this study, we applied a GAE framework to nationwide data from Korea spanning 2022 to 2023, integrating monthly air pollution measurements with emergency room visit data for 4 major cardiovascular and cerebrovascular disease outcomes: CA, MI, IS, and hemorrhagic stroke (HS). By constructing graphs based on correlation structures and training GAE models across different time points and regions, we aimed to uncover latent patterns in the pollutant–disease network. The objective of this study is to provide empirical evidence that can inform season- and region-specific public health strategies, enhance risk prediction, and support environmental health policy initiatives to reduce acute disease events triggered by air pollution exposure.

Methods

Ethics statement

This study used publicly available, de-identified secondary data obtained from national databases; therefore, it did not involve direct human participation or identifiable personal information. As a result, institutional review board approval and informed consent were not required under local regulations.

Study design

This research is a nationwide ecological analysis. The study is described in accordance with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) statement, available at https://www.strobe-statement.org/.

Setting

This study explored the association between air pollutant concentrations and emergency room visits for cardiovascular and cerebrovascular diseases in Korea from January 2022 to December 2023 (Supplement 1). To address the complex, nonlinear, and multivariate nature of environmental health data, a graph-based machine learning model (GAE) was applied to identify and visualize seasonal and regional association patterns.

Participants

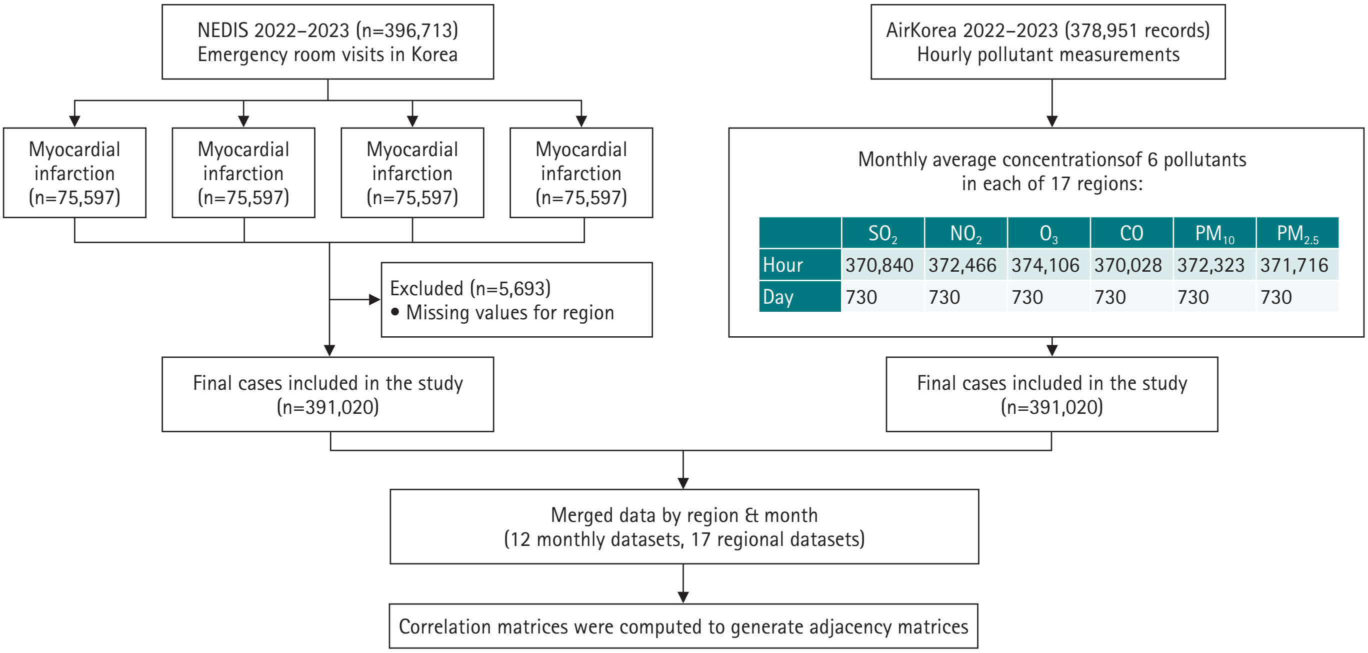
A total of 378,951 air pollution records and 396,713 emergency room visit cases were initially collected. After excluding 5,693 cases with missing or ambiguous regional information, the final analytic sample included 391,020 cases (Fig. 1, Supplement 1).

Variables

Exposure variables included 6 pollutants: SO_2, NO_2, ozone (O_3), carbon monoxide (CO), particulate matter ≤ 10 μm (PM_{10}), and particulate matter ≤ 2.5 μm ($PM_{2.5}$). Outcome variables comprised 4 major disease categories: CA, MI, IS, and HS.

Data sources

Two national datasets were utilized to examine the relationship between air pollution and emergency room visits for cardiovascular and cerebrovascular diseases. Air pollution data were obtained from the National Air Pollution Monitoring Network (AirKorea; https://www.airkorea.or.kr/web/pastSearch?pMENU_NO=123) and included hourly measurements of pollutants (Dataset 1). Disease data were sourced from the National Emergency Department Information System (NEDIS; https://www.e-gen.or.kr/


NEDIS 2022–2023 (n=396,713)
Emergency room visits in Korea
Myocardial infarction (n=75,597)
Myocardial infarction (n=75,597)
Myocardial infarction (n=75,597)
Myocardial infarction (n=75,597)
Excluded (n=5,693)
• Missing values for region
Final cases included in the study (n=391,020)
AirKorea 2022–2023 (378,951 records)
Hourly pollutant measurements
Monthly average concentrationsof 6 pollutants in each of 17 regions:
Final cases included in the study (n=391,020)
Merged data by region & month
(12 monthly datasets, 17 regional datasets)
Correlation matrices were computed to generate adjacency matrices

	SO_2	NO_2	O_3	CO	PM_{10}	$PM_{2.5}$
Hour	370,840	372,466	374,106	370,028	372,323	371,716
Day	730	730	730	730	730	730

Fig. 1. Overview of data sources, preprocessing, and analysis pipeline. This flowchart illustrates the data integration and preprocessing steps used in the study. Emergency room visit records (n=396,713) for 4 cardiovascular and cerebrovascular conditions—myocardial infarction (MI), hemorrhagic stroke (HS), ischemic stroke (IS), and cardiac arrest (CA)—were obtained from the National Emergency Department Information System (NEDIS) for 2022–2023. After excluding records with missing regional information (n=5,693), 391,020 cases were included in the final analysis. Simultaneously, hourly air pollutant data (n=378,951) from AirKorea were aggregated into monthly averages for 6 pollutants (SO_2, NO_2, O_3, CO, PM_{10}, and $PM_{2.5}$) across 17 regions. All variables were normalized using Min-Max scaling. The datasets were merged by region and month, resulting in 12 seasonal and 17 regional datasets. Correlation matrices were computed for each dataset to construct adjacency matrices, which were used as input for graph autoencoder modeling.

nemc/statistics_annual_report.do?brdclscd = 02), managed by the Ministry of Health and Welfare (Dataset 2). These datasets provided monthly counts of emergency room visits for the 4 major disease categories, recorded across 17 administrative regions.

Data preprocessing

Several preprocessing steps were conducted to enable integrated spatiotemporal analysis. First, hourly air pollution data were aggregated to monthly average concentrations for each of the 17 administrative regions. Monthly aggregation was adopted because emergency room visit counts from NEDIS were only available at a monthly frequency and at the provincial (si-do) level; pollutant measurements from individual monitoring stations were likewise aggregated to ensure consistency. Next, pollutant and disease datasets were merged by matching month and region, resulting in 2 structured input types: (1) time-based datasets for each 12-month period, used to explore seasonal variation, and (2) region-based datasets covering 17 provinces, used to assess spatial heterogeneity. Finally, all variables were normalized to a 0–1 scale using Min-Max scaling to standardize input values and facilitate stable model training (Fig 1, Supplement 2).

Measurement (graph construction)

The integrated dataset was transformed into graph structures for input into the GAE model (Supplements 2, 3). Each of the 10 key variables—comprising the 6 pollutants and 4 disease categories—was represented as a node in the graph. A Pearson correlation matrix was calculated to quantify the relationships between nodes. An edge was created between 2 nodes only if the absolute value of the Pearson correlation coefficient ($|\rho ij|$) was greater than or equal to 0.4, resulting in a sparse adjacency matrix (A). The corresponding node feature matrix (X) was defined such that n = 10 (number of nodes), and d represented the number of observations: 12 for time-series (monthly) analyses and 17 for regional (spatial) analyses.

Bias

As an ecological study, this analysis is subject to the risk of ecological fallacy, whereby associations observed at the population level may not apply to individuals. Potential confounding factors, such as meteorological conditions, socioeconomic status, and comorbidities, were not available in the dataset and could influence the observed associations.

Study size

All eligible records from national administrative databases for the defined period were analyzed; therefore, no sample size calculation was performed.

Statistical methods

The GAE model was implemented using the PyTorch Geometric library. The model architecture included an encoder with 2 graph convolutional layers (GCNConv) to compute node embeddings (Z), and a decoder that estimated edge probabilities between nodes by calculating the inner product of the embeddings, followed by a sigmoid activation function (σ). The loss function was defined as the binary cross-entropy between the original adjacency matrix (A) and the reconstructed matrix ($\hat{A}$). Model training was conducted for 100 epochs using the Adam optimizer with a learning rate of 0.01 (Supplements 2, 3). The resulting embedding vectors (Z) were visualized in 2-dimensional space to facilitate interpretation. Node positions were fixed based on the learned embeddings, and edge thickness was adjusted according to the predicted edge strength, visually reflecting the relative magnitude of pollutant–disease associations.

Two complementary analytical strategies were employed to explore the spatiotemporal dynamics of the pollutant–disease relationships. First, for the seasonal analysis, GAE models were trained individually for each month from January to December in both 2022 and 2023, allowing assessment of temporal variability. Second, for the regional analysis, the dataset was subdivided according to the 17 administrative regions, and separate GAE models were developed to examine spatial heterogeneity.

Additionally, a lag analysis was performed using pollutant concentrations from the previous month (t–1) to predict emergency room visits in the current month (t). The resulting graphs were exported as PNG images for visual inspection. In addition, the predicted edge strengths generated by the GAE and the corresponding Pearson correlation coefficients were compiled into Excel files (Microsoft Corp.) to enable comparative and supplementary quantitative analysis (Supplement 3).

Results

GAE models were applied to spatiotemporal datasets of air pollutants and emergency room visits for cardiovascular and cerebrovascular diseases across 17 administrative regions in Korea, spanning January 2022 to December 2023. The analysis revealed pronounced seasonal patterns in pollutant–disease associations, as well as distinct structural differences between urban and non-urban regions.

Seasonal variations

Using monthly GAE models, we constructed pollutant–disease networks to track how associations changed across seasons in 2022 (Fig. 2A–D, Supplements 4, 5). Each network visualizes the structural complexity and strength of connections between 6 major air pollutants—SO_2, NO_2, O_3, CO, PM_{10}, and $PM_{2.5}$—and 4 emergency room diagnoses: MI, IS, HS, and CA.

In January, the network displayed the densest connections of any month, with a high mean edge strength of 0.960 (Fig. 2A). Most nodes were highly interconnected, and NO_2 and O_3 emerged as central hubs, robustly linked not only to each disease outcome but also to the other pollutants, including SO_2, PM_{10}, $PM_{2.5}$, and CO. These widespread linkages suggest that exposure to multiple pollutants during winter imposes a cumulative burden on cardiovascular and cerebrovascular health. Notably, MI, IS, and CA all showed concurrent associations with nearly every pollutant, highlighting the particular vulnerability of these acute conditions to wintertime pollution.

By April, although the network remained complex—with the highest mean edge strength of the year at 0.967—the dominant pattern shifted (Fig. 2B). A highly integrated subnetwork formed among $PM_{2.5}$, PM_{10}, NO_2, and the 4 disease nodes, indicating that particulate matter and NO_2 were the primary contributors to emergency visits in spring. In contrast to the winter network, where O_3 was central, ozone became peripheral in spring, and both CO and SO_2 disconnected from the disease cluster. This reorganization points to seasonal specificity in pollutant effects. In particular, the edge between PM_{10} and CA, as well as between $PM_{2.5}$ and MI, was notably strengthened.

In August, the network showed a distinct bifurcation: pollutant nodes (NO_2, PM_{10}, $PM_{2.5}$, SO_2, CO, O_3) formed a tightly interconnected group, while disease nodes (MI, HS, IS, CA) clustered separately, with few or no direct links to pollutants (Fig. 2C). This configuration yielded the lowest mean correlation of the year, at 0.852. The reduced connectivity likely reflects decreased pollutant concentrations and enhanced atmospheric dispersion in summer, contributing to a temporary attenuation of acute health risks. Nevertheless, the strong internal cohesion among pollutants suggests that environmental exposure remained a concern, even in the absence of immediate disease associations.

By October, network connectivity rebounded, with a mean edge value of 0.905. NO_2 and O_3 reemerged as central nodes and

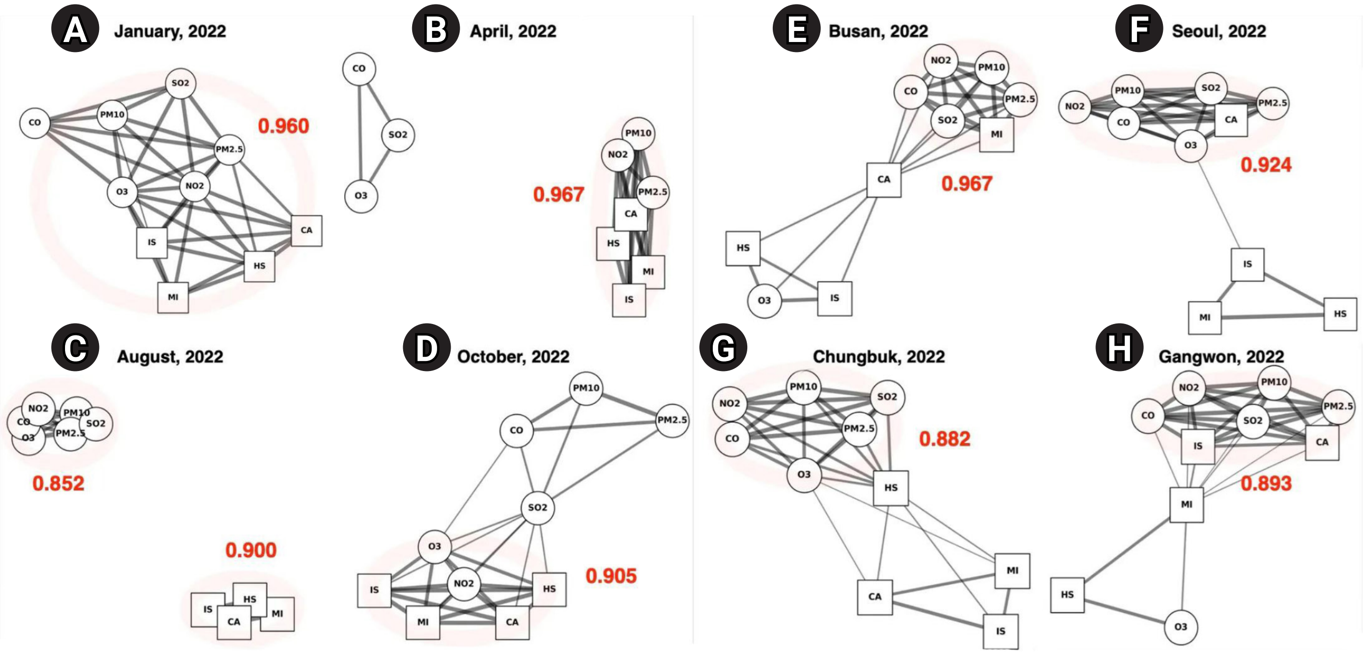


Fig. 2. Visualization of pollutant–disease networks by season and region (2022). This figure presents representative graph autoencoder (GAE) network outputs, showing seasonal (A–D) and regional (E–H) variations in the relationships between air pollutants and emergency room visit diseases in Korea during 2022. Circular nodes represent air pollutants (SO_2, NO_2, O_3, CO, PM_{10}, and $PM_{2.5}$), while square nodes represent disease categories: cardiac arrest (CA), myocardial infarction (MI), ischemic stroke (IS), and hemorrhagic stroke (HS). Edges indicate learned associations, with thicker lines representing stronger predicted relationships. Red numbers denote the average predicted edge strength (mean ≥0.5) for each network. Panels (A–D) show seasonal networks for January, April, August, and October 2022, respectively. Panels (E–H) display regional networks for Busan, Seoul, Chungbuk, and Gangwon in 2022. The visualizations highlight pronounced seasonal changes in network density and structure: winter (A) and spring (B) show high connectivity, while summer (C) is much sparser. Regional differences are also clear: urban areas like Busan (E) and Seoul (F) display more complex networks, whereas Chungbuk (G) and Gangwon (H) have less integrated structures.

re-established links with all 4 disease outcomes (Fig. 2D). In contrast to spring, CO and SO_2 were once again integrated into the disease–pollutant cluster, suggesting a broader pollutant profile influencing emergency room visits as temperatures cooled. A particularly dense connection between NO_2, MI, and CA was observed, closely resembling the winter pattern and indicating a resurgence of cardiovascular vulnerability. Notably, $PM_{2.5}$ and PM_{10} were less prominent in autumn than in spring, suggesting that gaseous pollutants once again played a leading role in disease associations during the colder transition period.

Regional differences

Regional-level analysis was performed using individual GAE models for each of Korea's 17 administrative regions, as shown in Fig. 2E–H and Supplements 6 and 7. This approach enabled detailed comparisons of network structures between urban and non-urban areas, revealing clear differences in the complexity and strength of pollutant–disease associations.

Urban regions such as Busan and Seoul exhibited highly connected and densely structured networks. In Busan, the network had the highest regional correlation coefficient (0.967), with CA centrally positioned and strongly linked to several pollutants, including NO_2, SO_2, PM_{10}, and $PM_{2.5}$. MI also showed robust connections, particularly with fine particulate matter, suggesting that cardiovascular risk in this metropolitan area is influenced by a broad spectrum of air pollutants (Fig. 2E). In Seoul, a similarly complex subnetwork formed among gaseous pollutants (NO_2, CO, O_3, SO_2), PMs, and CA. IS, MI, and HS were separated from this main cluster, while the remaining nodes interlinked to form an overall dense structure, with a correlation coefficient of 0.924 (Fig. 2F). Non-urban areas presented a contrasting profile. In Chungbuk, the network was moderately sparse (0.882), with pollutant nodes densely interconnected but showing fewer links to disease nodes. HS had limited connections, while MI and IS remained more peripheral (Fig. 2G). In Gangwon, the structure was slightly denser (0.893) than in Chungbuk but still showed some disconnection between certain diseases and pollutants (Fig. 2H). MI emerged as a bridge between the pollutant cluster—especially

SO_2 and NO_2—and other health outcomes, while O_3 and HS were more isolated. These findings indicate that non-metropolitan regions, while not devoid of pollutant–disease interactions, experience weaker or less complex exposures compared to urban centers, likely reflecting differences in emission sources, population density, and environmental conditions.

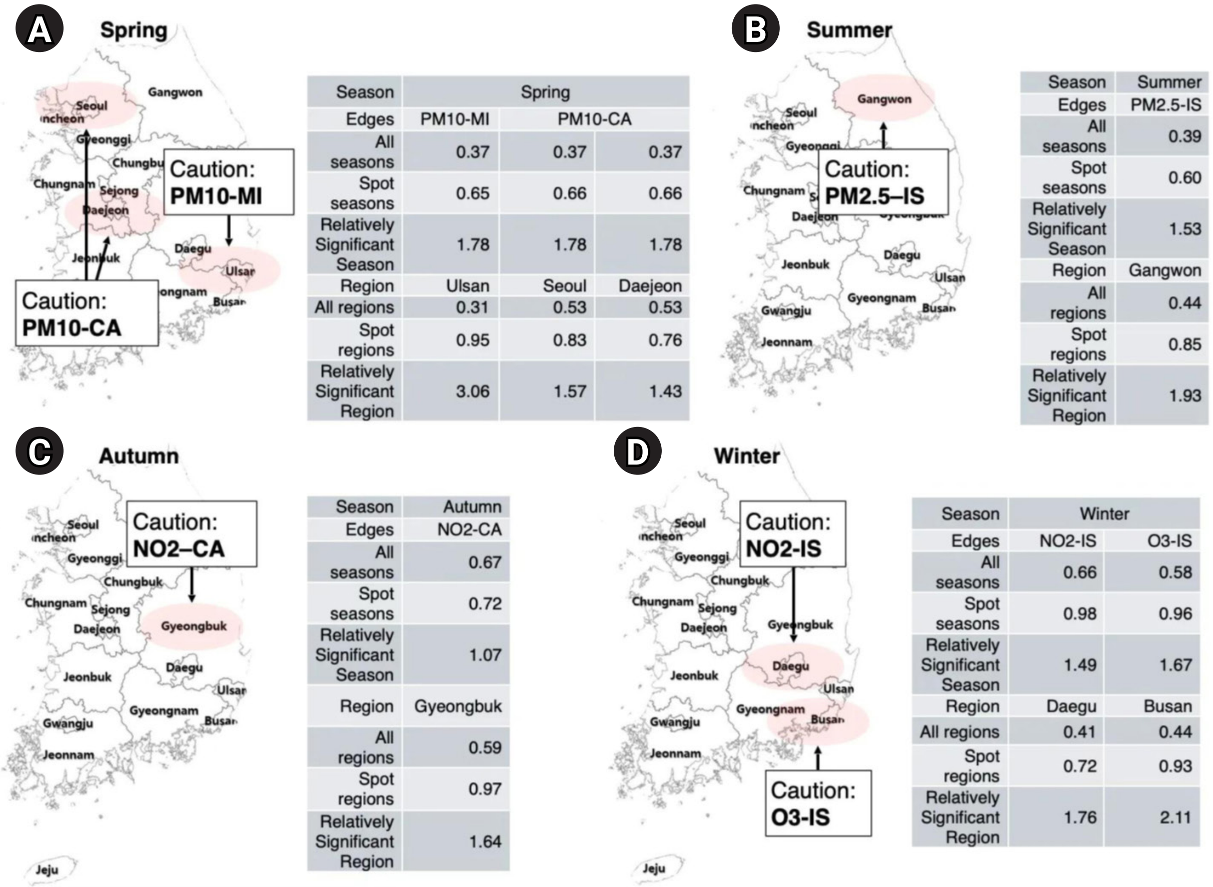
Lag analysis

Supplements 8–11 present lagged GAE networks in which pollutant concentrations from the previous month (t–1) are used to predict emergency room visits in the current month (t). The lagged networks exhibit seasonal patterns similar to those found in the contemporaneous analysis: in winter, strong NO_2–O_3 and disease node linkages persist, while in summer, overall network connectivity remains attenuated. These parallel findings confirm that the monthly co-occurrence patterns identified in the primary analysis are robust to a 1-month temporal shift.

Discussion

Key results

This study applied a GAE model to evaluate the spatiotemporal associations between air pollutants and emergency room visits for cardiovascular and cerebrovascular diseases across Korea during 2022–2023. The findings revealed pronounced seasonal and regional variations (Fig. 3A–D). Strong associations were observed



Season	Spring		
Edges	PM10-MI	PM10-CA	
All seasons	0.37	0.37	0.37
Spot seasons	0.65	0.66	0.66
Relatively Significant Season	1.78	1.78	1.78
Region	Ulsan	Seoul	Daejeon
All regions	0.31	0.53	0.53
Spot regions	0.95	0.83	0.76
Relatively Significant Region	3.06	1.57	1.43

Season	Summer
Edges	PM2.5-IS
All seasons	0.39
Spot seasons	0.60
Relatively Significant Season	1.53
Region	Gangwon
All regions	0.44
Spot regions	0.85
Relatively Significant Region	1.93

Season	Autumn
Edges	NO2-CA
All seasons	0.67
Spot seasons	0.72
Relatively Significant Season	1.07
Region	Gyeongbuk
All regions	0.59
Spot regions	0.97
Relatively Significant Region	1.64

Season	Winter	
Edges	NO2-IS	O3-IS
All seasons	0.66	0.58
Spot seasons	0.98	0.96
Relatively Significant Season	1.49	1.67
Region	Daegu	Busan
All regions	0.41	0.44
Spot regions	0.72	0.93
Relatively Significant Region	1.76	2.11

Fig. 3. Spatiotemporal hotspots of strong pollutant–disease associations by season. This figure highlights region- and season-specific pollutant–disease pairs with relatively strong associations based on graph autoencoder predictions. Each panel presents a map of South Korea with red-shaded areas indicating regions with significantly elevated edge strengths. Accompanying tables summarize the mean edge values across all seasons, within the highlighted season, and for specific regions, along with their relative significance. (A) Spring: PM_{10} is strongly associated with myocardial infarction (MI) in Ulsan and with cardiac arrest (CA) in Seoul and Daejeon. These associations are most prominent in spring, showing more than 3-fold increases in relative significance in the hotspot regions. (B) Summer: Gangwon is identified as a hotspot for $PM_{2.5}$–ischemic stroke (IS) associations, with a relative increase in edge strength during summer. (C) Autumn: NO_2 exhibits a strong seasonal link with cardiac arrest in Gyeongbuk, indicating increased vulnerability in this region during autumn. (D) Winter: NO_2–IS and O_3–IS associations are especially elevated in Daegu and Busan, respectively, with winter showing the highest relative significance for these edges (up to 2.11-fold).

in winter between NO_2/O_3 and IS, in spring between particulate matter (PM_{10} and $PM_{2.5}$) and MI or CA, and in autumn between NO_2 and CA. Urban regions consistently displayed more complex pollutant–disease networks than non-urban areas, highlighting differential exposure and vulnerability [14].

Interpretation

These seasonal shifts in network structure reflect both bulk concentration changes and mechanistic compositional differences. In winter, elevated NO_2 and O_3 levels—driven by increased residential heating emissions and atmospheric stagnation—contrast with higher PM mass in spring, attributable to dust storms and intensified outdoor activity [15,16]. Source-apportionment studies further indicate that secondary inorganic aerosols (sulfate, nitrate, ammonium) make up approximately 64% of $PM_{2.5}$ in winter (compared to a 49% annual average), with a sulfate-to-nitrate mass ratio of around 1.5:1. In spring (March–May), crustal dust contributions increase from 8% to 12%, and the combined sulfate plus nitrate fraction rises from 39% to 45% [17,18]. The attenuated associations in summer could be due to improved pollutant dispersion and increased indoor activities, such as staying in air-conditioned environments, which lower personal exposure; however, persistent risks in regions like Gangwon suggest localized vulnerabilities [15,19-22].

The regional differences, particularly the denser connections in metropolitan centers, underscore health disparities resulting from urbanization [23,24]. These results fulfill the study's objectives by identifying critical periods and locations for targeted air quality and health interventions. The GAE approach provided a robust framework for detecting nonlinear, multivariate relationships that conventional statistical models may miss [25].

Given these spatiotemporal dynamics, our findings support the need for season-specific and region-specific public health strategies. In winter (Fig. 3D), strict management of NO_2 and O_3—along with focused health monitoring of at-risk populations in major cities such as Busan, Daegu, and the Seoul Metropolitan Area—is essential [26,27]. In spring, effective control of particulate matter (PM_{10} and $PM_{2.5}$) and improvements in indoor air quality in urban centers like Seoul, Daejeon, and Ulsan may help reduce emergency cardiovascular incidents [28] (Fig. 3A). In autumn, targeted NO_2 reduction strategies should be prioritized in non-metropolitan areas, such as Gyeongbuk and Chungbuk, to mitigate risks of stroke and CA [29] (Fig. 3C). During summer, inland regions such as Gangwon may benefit from region-specific surveillance systems and risk assessment models that incorporate meteorological factors, which could explain sustained vulnerability despite improved average air quality [30] (Fig. 3B).

Comparison with previous studies

Our findings are consistent with previous research documenting stronger PM–stroke associations under extreme winter conditions in Seoul [31] and springtime soil-dust peaks in Incheon, as well as autumn industrial/coal impacts in Daegu [32]. These studies document seasonal and spatial heterogeneity in pollution–health relationships. However, prior research has generally focused on single pollutants or individual cities, whereas our GAE-based network analysis integrates multi-pollutant, multi-region time-series data to reveal how and when pollutant–disease clusters form across Korea [33]. This network approach uncovers interconnected risk modules—such as winter NO_2/O_3–stroke and spring PM–MI clusters—and tracks their spatiotemporal evolution, offering dynamic insights and identifying periods and subregional vulnerabilities that traditional models may not capture [34]. Notably, our network-level findings recent policy evaluations in Seoul showing that long-term air quality improvements correspond with reductions in cardiovascular morbidity, reinforcing the real-world value of targeted interventions [35].

Limitations

Several limitations should be acknowledged. First, the analysis was restricted to a 2-year period, which may not capture longer-term trends or lagged effects of pollution exposure. Second, health data were limited to emergency room visits, which may underestimate the overall disease burden. Third, the study used only monthly aggregated NEDIS emergency room visit data, which precluded direct assessment of daily or weekly acute exposure–effect relationships. Although monthly aggregation reduces noise and computational burden, it may mask short-lag effects. Fourth, the analysis was conducted at the provincial (si-do) level (17 regions), potentially missing intra-regional heterogeneity (e.g., variation in pollution and emergency room visit patterns within Seoul). Fifth, although the GAE model captured key pollutant–disease associations, it did not adjust for meteorological variables (such as temperature or humidity), individual-level risk factors, or socioeconomic status, all of which could confound or modify the observed associations. The exclusion of these covariates may have led to attenuation or inflation of some edge strengths, depending on region and season.

Generalizability

Despite these constraints, the study's national scope and use of standardized administrative data enhance its generalizability within Korea. The methodological framework—particularly the GAE modeling—can be adapted to other countries with similar air pollution and health data systems. The visual and analytical insights derived from this model are valuable for healthcare policymakers, urban planners, and environmental health professionals aiming to develop locally tailored public health interventions.

Suggestions for further studies

Future studies should expand the observation period and include a broader array of health outcomes, such as hospitalizations and mortality. Incorporating meteorological, behavioral, and socioeconomic data would improve model precision and policy relevance. Further, investigating lagged effects of pollutant exposure and applying explainable AI methods to graph models may offer deeper insight into causal mechanisms. Ultimately, the integration of network science and public health surveillance offers a powerful approach for disease prevention and environmental risk management in increasingly urbanized societies.

High-resolution si/gun/gu-level graph analyses (corresponding to cities, counties, and districts), integrating inter-district movement matrices with moving-average pollutant exposure metrics, could trace subregional pollution–disease diffusion in a time-series framework, supporting improved causal inference and more targeted interventions.

Conclusion

This study investigated the spatiotemporal dynamics of emergency room visits associated with air pollution exposure in Korea by applying a GAE model to national datasets from 2022 to 2023. The analysis identified distinct seasonal and regional patterns in the associations between 6 major air pollutants and 4 categories of cardiovascular and cerebrovascular diseases.

We found that the strength and structure of pollutant–disease relationships varied significantly by season and urbanization level. Notably, NO_2 and O_3 played dominant roles in disease associations during winter and autumn, while particulate matter (PM_{10} and $PM_{2.5}$) showed strong links to cardiac events in spring. In contrast, summer was characterized by overall weaker associations, though inland areas such as Gangwon continued to exhibit localized vulnerability. Urban regions consistently exhibited denser and more complex pollutant–disease networks compared to non-urban areas. These findings support the initial research aim of revealing the complex, multivariate relationships between environmental exposures and acute health outcomes across time and space. The use of GAE modeling enabled us to capture and visualize these nonlinear associations more effectively than conventional statistical methods.

From a medical and public health perspective, this study highlights the importance of season-specific and region-specific strategies for air pollution management and disease prevention. By pinpointing when and where pollutant–disease associations are most pronounced, our results can inform targeted interventions, risk communication, and policy planning to reduce preventable emergency health events and improve cardiovascular and cerebrovascular health nationwide.

ORCID

Sohee Wang: https://orcid.org/0009-0007-4324-6034
Seungpil Jeong: https://orcid.org/0000-0002-1766-1425
Eunhee Ha: https://orcid.org/0000-0002-4224-3858

Authors' contribution

Conceptualization: EH. Data curation: SJ. Methodology/formal analysis/validation: SW, SJ. Project administration: EH. Funding acquisition: none. Writing–original draft: SW. Writing–review & editing: SW, SJ, EH.

Conflict of interest

No potential conflict of interest relevant to this article was reported.

Funding

None.

Data availability

The datasets analyzed in this study are publicly available from the following sources: https://doi.org/10.7910/DVN/ZY7MVM

Dataset 1. Hourly air pollution measurement data (SO_2, NO_2, O_3, CO, PM_{10}, $PM_{2.5}$) from 17 administrative regions in Korea (2022–2023) can be accessed via the AirKorea Final Confirmation Measurement Data Portal: (https://www.airkorea.or.kr/web/pastSearch?pMENU_NO = 123).

Dataset 2. Emergency room visit records related to cardiovascular and cerebrovascular diseases (2022–2023) are available from the NEDIS Emergency Medical Portal Statistical Yearbook: (https://www.e-gen.or.kr/nemc/statistics_annual_report.do?brdclscd = 02).

All data used in this study are publicly accessible, and no proprietary or restricted datasets were used. The code used for analysis is available on GitHub: (https://github.com/seungpil720/sohee).

Acknowledgments

None.

Supplementary materials

Supplementary files are available from https://doi.org/10.7910/

DVN/ZY7MVM

Supplement 1. Temporal trends in daily air pollutant concentrations and monthly disease incidence (2022–2023).

Supplement 2. Overall workflow of data preprocessing and graph autoencoder (GAE) modeling.

Supplement 3. Mathematical notation of input representation, edge construction, and loss function in the graph autoencoder.

Supplement 4. Monthly graph autoencoder (GAE) networks depicting pollutant–disease associations in 2022.

Supplement 5. Monthly graph autoencoder (GAE) networks depicting pollutant–disease associations in 2023.

Supplement 6. Regional graph autoencoder (GAE) networks depicting pollutant–disease associations in 2022.

Supplement 7. Regional graph autoencoder (GAE) networks depicting pollutant–disease associations in 2023.

Supplement 8. Monthly graph autoencoder (GAE) Networks depicting lagged pollutant–disease associations in 2022.

Supplement 9. Monthly graph autoencoder (GAE) networks depicting lagged pollutant–disease associations in 2023.

Supplement 10. Regional graph autoencoder (GAE) networks depicting lagged pollutant–disease associations in 2022.

Supplement 11. Monthly graph autoencoder (GAE) networks depicting lagged pollutant–disease associations in 2023.

References

1. Lee HH, Cho SM, Lee H, Baek J, Bae JH, Chung WJ, Kim HC. Korea heart disease fact sheet 2020: analysis of nationwide data. Korean Circ J 2021;51:495-503. https://doi.org/10.4070/kcj.2021.0097
2. de Bont J, Jaganathan S, Dahlquist M, Persson A, Stafoggia M, Ljungman P. Ambient air pollution and cardiovascular diseases: an umbrella review of systematic reviews and meta-analyses. J Intern Med 2022;291:779-800. https://doi.org/10.1111/joim.13467
3. Toubasi A, Al-Sayegh TN. Short-term exposure to air pollution and ischemic stroke: a systematic review and meta-analysis. Neurology 2023;101:e1922-e1932. https://doi.org/10.1212/WNL.0000000000207856
4. Zou L, Zong Q, Fu W, Zhang Z, Xu H, Yan S, Mao J, Zhang Y, Cao S, Lv C. Long-term exposure to ambient air pollution and myocardial infarction: a systematic review and meta-analysis. Front Med (Lausanne) 2021;8:616355. https://doi.org/10.3389/fmed.2021.616355
5. Pan C, Xu C, Zheng J, Song R, Lv C, Zhang G, Tan H, Ma Y, Zhu Y, Han X, Li C, Yan S, Zheng W, Wang C, Zhang J, Bian Y, Ma J, Cheng K, Liu R, Hou Y, Chen Q, Zhao X, McNally B, Chen R, Kan H, Meng X, Chen Y, Xu F. Fine and coarse particulate air pollution and out-of-hospital cardiac arrest onset: a nationwide case-crossover study in China. J Hazard Mater 2023;457:131829. https://doi.org/10.1016/j.jhazmat.2023.131829
6. Kim H, Kim J, Kim S, Kang SH, Kim HJ, Kim H, Heo J, Yi SM, Kim K, Youn TJ, Chae IH. Cardiovascular effects of long-term exposure to air pollution: a population-based study with 900 845 person-years of follow-up. J Am Heart Assoc 2017;6:e007170. https://doi.org/10.1161/JAHA.117.007170
7. Cesaroni G, Forastiere F, Stafoggia M, Andersen ZJ, Badaloni C, Beelen R, Caracciolo B, de Faire U, Erbel R, Eriksen KT, Fratiglioni L, Galassi C, Hampel R, Heier M, Hennig F, Hilding A, Hoffmann B, Houthuijs D, Jöckel KH, Korek M, Lanki T, Leander K, Magnusson PK, Migliore E, Ostenson CG, Overvad K, Pedersen NL, J JP, Penell J, Pershagen G, Pyko A, Raaschou-Nielsen O, Ranzi A, Ricceri F, Sacerdote C, Salomaa V, Swart W, Turunen AW, Vineis P, Weinmayr G, Wolf K, de Hoogh K, Hoek G, Brunekreef B, Peters A. Long term exposure to ambient air pollution and incidence of acute coronary events: prospective cohort study and meta-analysis in 11 European cohorts from the ESCAPE Project. BMJ 2014;348:f7412. https://doi.org/10.1136/bmj.f7412
8. Mustafic H, Jabre P, Caussin C, Murad MH, Escolano S, Tafflet M, Perier MC, Marijon E, Vernerey D, Empana JP, Jouven X. Main air pollutants and myocardial infarction: a systematic review and meta-analysis. JAMA 2012;307:713-721. https://doi.org/10.1001/jama.2012.126
9. Billionnet C, Sherrill D, Annesi-Maesano I. Estimating the health effects of exposure to multi-pollutant mixture. Ann Epidemiol 2012;22:126-141. https://doi.org/10.1016/j.annepidem.2011.11.004
10. Hooper LG, Kaufman JD. Ambient air pollution and clinical implications for susceptible populations. Ann Am Thorac Soc 2018;15:S64-S68. https://doi.org/10.1513/AnnalsATS.201707-574MG
11. Dominski FH, Lorenzetti Branco JH, Buonanno G, Stabile L, Gameiro da Silva M, Andrade A. Effects of air pollution on health: a mapping review of systematic reviews and meta-analyses. Environ Res 2021;201:111487. https://doi.org/10.1016/j.envres.2021.111487
12. Bazyar J, Pourvakhshoori N, Khankeh H, Farrokhi M, Delshad V, Rajabi E. A comprehensive evaluation of the association between ambient air pollution and adverse health outcomes of major organ systems: a systematic review with a worldwide approach. Environ Sci Pollut Res Int 2019;26:12648-12661. https://doi.org/10.1007/s11356-019-04874-z
13. Kipf TN, Welling M. Variational graph auto-encoders. arXiv

[Preprint] 2016 Nov 21. https://doi.org/10.48550/arXiv.1611.07308

14. Kim OJ, Kim SY. Regional difference in the association between long-term PM and cardiovascular disease incidence and potential determinants of the difference. Geohealth 2025;9:e2024GH001245. https://doi.org/10.1029/2024GH001245
15. Cichowicz R, Wielgosinski G, Fetter W. Dispersion of atmospheric air pollution in summer and winter season. Environ Monit Assess 2017;189:605. https://doi.org/10.1007/s10661-017-6319-2
16. Olvera Alvarez HA, Myers OB, Weigel M, Armijos RX. The value of using seasonality and meteorological variables to model intra-urban PM variation. Atmos Environ (1994) 2018;182:1-8. https://doi.org/10.1016/j.atmosenv.2018.03.007
17. Kim H, Zhang Q, Bae GN, Kim JY, Lee SB. Sources and atmospheric processing of winter aerosols in Seoul, Korea: insights from real-time measurements using a high-resolution aerosol mass spectrometer. Atmos Chem Phys 2017;17:2009-2033. https://doi.org/10.5194/acp-17-2009-2017
18. Park EH, Heo J, Kim H, Yi SM. Long term trends of chemical constituents and source contributions of PM in Seoul. Chemosphere 2020;251:126371. https://doi.org/10.1016/j.chemosphere.2020.126371
19. Li Q, Zhang H, Jin X, Cai X, Song Y. Mechanism of haze pollution in summer and its difference with winter in the North China Plain. Sci Total Environ 2022;806:150625. https://doi.org/10.1016/j.scitotenv.2021.150625
20. Jung SH, Baek SH, Park SY, Lee CM, Lee JI. Regional differences in PM chemical composition and inhalation risk assessment: a case study of Seoul, Incheon, and Wonju. Toxics 2025;13:240. https://doi.org/10.3390/toxics13040240
21. Lee UJ, Kim MJ, Kim EJ, Lee DW, Lee SD. Spatial distribution characteristics and analysis of PM2. 5 in South Korea: a geographically weighted regression analysis. Atmosphere 2024;15:69. https://doi.org/10.3390/atmos15010069
22. Bell ML, Ebisu K, Peng RD, Dominici F. Adverse health effects of particulate air pollution: modification by air conditioning. Epidemiology 2009;20:682-686. https://doi.org/10.1097/EDE.0b013e3181aba749
23. Wang Q. Urbanization and global health: the role of air pollution. Iran J Public Health 2018;47:1644-1652.
24. Samet JM, White RH. Urban air pollution, health, and equity. J Epidemiol Community Health 2004;58:3-5. https://doi.org/10.1136/jech.58.1.3
25. Liu M, Zhang Z, Dunson DB. Graph auto-encoding brain networks with applications to analyzing large-scale brain imaging datasets. Neuroimage 2021;245:118750. https://doi.org/10.1016/j.neuroimage.2021.118750
26. Colombi NK, Jacob DJ, Yang LH, Zhai S, Shah V, Grange SK, Yantosca RM, Kim S, Liao H. Why is ozone in South Korea and the Seoul metropolitan area so high and increasing? Atmos Chem Phys 2023;23:4031-4044. https://doi.org/10.5194/acp-23-4031-2023
27. Bae HJ, Shin YS, Park JG, Pak H. Interactive effects between ozone and temperature on respiratory admissions in Korea. ISEE Conf Abstr 2013;2013:5102. https://doi.org/10.1289/isee.2013.P-1-12-19
28. Kang SH, Heo J, Oh IY, Kim J, Lim WH, Cho Y, Choi EK, Yi SM, Do Shin S, Kim H, Oh S. Ambient air pollution and out-of-hospital cardiac arrest. Int J Cardiol 2016;203:1086-1092. https://doi.org/10.1016/j.ijcard.2015.11.100
29. Song J, Lim Y, Ko I, Kim JY, Kim DK. Association between air pollutants and initial hospital admission for ischemic stroke in Korea from 2002 to 2013. J Stroke Cerebrovasc Dis 2021;30:106080. https://doi.org/10.1016/j.jstrokecerebrovasdis.2021.106080
30. Yeo MJ, Kim YP. Long-term trends and affecting factors in the concentrations of criteria air pollutants in South Korea. J Environ Manage 2022;317:115458. https://doi.org/10.1016/j.jenvman.2022.115458
31. Park SK. Seasonal variations of fine particulate matter and mortality rate in Seoul, Korea with a focus on the short-term impact of meteorological extremes on human health. Atmosphere 2021;12:151. https://doi.org/10.3390/atmos12020151
32. Choi E, Yi SM, Lee YS, Jo H, Baek SO, Heo JB. Sources of airborne particulate matter-bound metals and spatial-seasonal variability of health risk potentials in four large cities, South Korea. Environ Sci Pollut Res Int 2022;29:28359-28374. https://doi.org/10.1007/s11356-021-18445-8
33. Testa A, Biondi-Zoccai G, Anticoli S, Pezzella FR, Mangiardi M, DI Giosa A, Marchegiani G, Frati G, Sciarretta S, Perrotta A, Peruzzi M, Cavarretta E, Gaspardone A, Mariano E, Federici M, Montone RA, Dei Giudici A, Versaci B, Versaci F. Cluster analysis of weather and pollution features and its role in predicting acute cardiac or cerebrovascular events. Minerva Med 2022;113:825-832. https://doi.org/10.23736/S0026-4806.22.08036-3
34. Dias D, Tchepel O. Spatial and temporal dynamics in air pollution exposure assessment. Int J Environ Res Public Health 2018;15:558. https://doi.org/10.3390/ijerph15030558
35. Han C, Lim YH, Yorifuji T, Hong YC. Air quality management policy and reduced mortality rates in Seoul Metropolitan Area: a quasi-experimental study. Environ Int 2018;121:600-609. https://doi.org/10.1016/j.envint.2018.09.047

eISSN: 2234-2591
Open Access

Original article

Ewha Med J 2025;48(3):e44
https://doi.org/10.12771/emj.2025.00486

The Ewha Medical Journal

Automated early detection of androgenetic alopecia using deep learning on trichoscopic images from a Korean cohort: a retrospective model development and validation study

Min Jung Suh[1], Sohyun Ahn[2*], Ji Yeon Byun[3]

[1]*Ewha Womans University College of Medicine, Seoul, Korea*
[2]*Ewha Medical Research Institute, Ewha Womans University College of Medicine, Seoul, Korea*
[3]*Department of Dermatology, Ewha Womans University College of Medicine, Seoul, Korea*

Purpose: This study developed and validated a deep learning model for the automated early detection of androgenetic alopecia (AGA) using trichoscopic images, and evaluated the model's diagnostic performance in a Korean clinical cohort.
Methods: We conducted a retrospective observational study using 318 trichoscopic scalp images labeled by board-certified dermatologists according to the Basic and Specific (BASP) system, collected at Ewha Womans University Medical Center between July 2018 and January 2024. The images were categorized as BASP 0 (no hair loss) or BASP 1–3 (early-stage hair loss). A ResNet-18 convolutional neural network, pretrained on ImageNet, was fine-tuned for binary classification. Internal validation was performed using stratified 5-fold cross-validation, and external validation was conducted through ensemble soft voting on a separate hold-out test set of 20 images. Model performance was measured by accuracy, precision, recall, F1-score, and area under the curve (AUC), with 95% confidence intervals (CIs) calculated for hold-out accuracy.
Results: Internal validation revealed robust model performance, with 4 out of 5 folds achieving an accuracy above 0.90 and an AUC above 0.93. In external validation on the hold-out test set, the ensemble model achieved an accuracy of 0.90 (95% CI, 0.77–1.03) and an AUC of 0.97, with perfect recall for early-stage hair loss. No missing data were present, and the model demonstrated stable convergence without requiring data augmentation.
Conclusion: This model demonstrated high accuracy and generalizability for detecting early-stage AGA from trichoscopic images, supporting its potential utility as a screening tool in clinical and teledermatology settings.

Keywords: Alopecia; Computer neural networks; Scalp; Deep learning; Dermatologists

Introduction

Background/rationale

Hair loss, especially androgenetic alopecia (AGA), is a common dermatological condition that has a considerable impact on patients' quality of life. Early detection is essential, both for initiating timely treatment and for preventing further progression during the subtle and potentially reversible stages of the disease [1]. In clinical settings, the Basic and Specific (BASP) classification system is widely used to assess the severity of hair loss, systematically categorizing frontal and vertex scalp patterns into structured scores [2]. However, BASP scoring is based on manual visual assessment, which introduces subjectivity and variability between observers.

To overcome these limitations, deep learning–based approaches have increasingly been explored in dermatology, providing automated and objective tools for image-based diagnosis [3]. Convolutional neural networks (CNNs), in particular, have shown strong performance in medical imaging tasks [4], including trichoscopic image analysis [5]. Building on this foundation, our study aimed to develop and validate a deep learning–based classification model capable of distinguishing BASP 0 (no hair loss) from BASP 1–3 (early-stage hair loss) directly from scalp images, with the goal of improving diagnostic reproducibility and stan-

*Corresponding email: mpsohyun@ewha.ac.kr
Received: June 23, 2025 Revised: July 15, 2025 Accepted: July 16, 2025

dardization.

This study was specifically designed to address common challenges in medical image analysis, such as small dataset size and class imbalance. To evaluate model performance, we employed a 2-tier experimental strategy. First, we performed internal validation using stratified 5-fold cross-validation across the entire dataset to assess training stability and to identify optimal training configurations [6]. Next, informed by these findings, we conducted external validation with a hold-out test set using ensemble voting, thereby simulating real-world application on previously unseen images [7]. This sequential approach allowed us to evaluate both model training dynamics and real-world generalizability.

To support these experiments, we redefined BASP labels into binary categories (BASP 0 vs. BASP 1–3), implemented class-preserving validation through stratified sampling, and selected ResNet-18, a lightweight yet effective CNN architecture known for its balance of performance and computational efficiency in small-to medium-sized datasets.

Objectives

The aim of this study was to evaluate the model's diagnostic accuracy, generalizability, and clinical utility using both internal cross-validation and external hold-out testing, providing evidence for its potential application in dermatological screening. Additionally, we sought to address class imbalance through stratified sampling, to assess the feasibility of binary BASP classification, and to demonstrate the use of a ResNet-18 CNN for the automated assessment of early-stage AGA.

Methods

Ethics statement

This study was approved by the Institutional Review Board (IRB) of Ewha Womans University Medical Center (IRB no., EUMC 2025-01-037). The requirement for informed consent was waived due to the retrospective nature of the study and the use of de-identified image data.

Study design

This was a retrospective observational study aimed at developing a deep learning model for classifying hair loss severity based on the BASP system.

Settings

A total of 318 trichoscopic images were collected from patients visiting the Department of Dermatology at Ewha Womans University Medical Center (Seoul, Republic of Korea) between July 7, 2018, and January 31, 2024. All images were acquired using the DermLite DL Cam Photo dermoscopy system (3Gen Inc.) and were captured from the frontal and vertex scalp regions during routine clinical assessments. Images of the occipital region, though used by dermatologists for clinical comparison during BASP scoring, were excluded from both model training and evaluation.

Participants

Eligible participants included patients aged 15 to 84 years who presented with concerns regarding hair loss. No additional inclusion or exclusion criteria were applied beyond clinical presentation, and all trichoscopic images with valid BASP annotations were included in the study. Labeling was performed by board-certified dermatologists using the BASP classification system, resulting in 151 images labeled as BASP 0 and 167 images labeled as BASP 1, 2, or 3 (Table 1). There were no missing data in the final dataset used for model development and evaluation.

Variables

The primary outcome variable was binary classification of hair loss severity, defined as class 0 for BASP 0 and class 1 for BASP 1–3. This binary categorization was derived from the original 4-class BASP labels (BASP 0, 1, 2, and 3), which were assigned by dermatologists.

Image preprocessing and model configuration

Each trichoscopic image was paired with its corresponding BASP score using a consistent filename-label mapping system, enabling the model to learn the association between image features and hair loss severity. Images were resized to 224 × 224 pixels and normalized to a mean of 0.5 and standard deviation of 0.5 for each RGB channel.

A ResNet-18 CNN pretrained on ImageNet was used as the backbone. The final fully connected layer was replaced with a 2-unit output layer for binary classification. Model training was performed using the Adam optimizer (learning rate = 0.001), a batch size of 128, and the cross-entropy loss function. No data augmentation or early stopping strategies were used. Each model was trained for a fixed 100 epochs.

Table 1. Distribution of original data

BASP label	0	1	2	3	Total
No. of images	151	109	47	11	318

BASP, Basic and Specific.

Validation test

Internal validation: Stratified 5-fold cross-validation

The entire dataset (n = 318) underwent stratified 5-fold cross-validation to ensure class balance within each fold. In every fold, 80% of the data were used for training and 20% for evaluation. The model with the highest validation accuracy over the 100 epochs was selected for reporting performance.

External validation: Hold-out ensemble testing

To evaluate generalizability, a hold-out test set of 20 images was created by randomly sampling 10 images from BASP 0 and 10 from BASP 1–3, ensuring class balance. The BASP 1–3 subset consisted of 4 BASP 1, 3 BASP 2, and 3 BASP 3 images, reflecting the distribution of hair loss stages (Table 2). These 20 images were completely excluded from model training and validation processes.

Model training and prediction

The remaining 298 images were used to train 5 ResNet-18 models via stratified 5-fold cross-validation. Each model was then applied to the hold-out set. Final predictions were determined by ensemble soft voting, where the average class probabilities from the 5 models were combined to determine the predicted label. This ensemble approach was intended to simulate a real-world diagnostic scenario and enhance robustness in performance estimation.

Bias mitigation strategies

To address potential sources of bias in this small and imbalanced dataset, stratified k-fold cross-validation was used to ensure that all original samples were included in both training and validation while maintaining class distribution across folds. This approach mitigated selection bias and maximized data utility. Data augmentation was intentionally excluded to avoid introducing artificial variability. Additionally, ensemble prediction via soft voting across 5-fold-specific models was used in the hold-out test phase to reduce model variance and improve generalizability.

Study size

In total, 318 images were analyzed. No a priori sample size calculation was performed; instead, all eligible labeled images from the institutional database were used to reflect real-world clinical data availability. Post hoc 95% confidence intervals (CIs) were calculated for model accuracy in the hold-out set, based on the primary endpoint of binary classification performance.

Table 2. Dataset configuration for hold-out ensemble voting test

	Class 0	Class 1		
BASP label	0	1	2	3
Hold-out test set	10	4	3	3

BASP, Basic and Specific.

Evaluation metrics

Performance on the hold-out set was assessed using ensemble accuracy, confusion matrix, receiver operating characteristic (ROC) curve analysis, area under the curve (AUC), and the 95% CI for accuracy. Classification metrics included accuracy, precision, recall, F1-score, and AUC. For the hold-out evaluation, a 95% CI for accuracy was computed using the Wald method.

Statistical methods

All statistical analyses and model training were performed using Python ver. 3.9 (https://www.python.org/) and PyTorch ver. 1.12 (Meta) in the Google Colab environment. Image preprocessing, model definition, training, and evaluation were implemented using in-house PyTorch-based scripts. Visualization of results, including ROC curves and confusion matrices, was conducted using Matplotlib ver. 3.7 (Hunter). No statistical hypothesis testing (such as P-values) was conducted, as the focus was on classification performance and generalizability rather than group comparisons. Python code is available in Supplement 1.

Results

Participants

A total of 318 trichoscopic scalp images were included for binary classification. Of these, 159 images were labeled as BASP 0 (no hair loss, class 0) and 159 images were labeled as BASP 1, 2, or 3 (early hair loss, class 1). No data were excluded, and all labeled images were used in both model training and evaluation.

Internal validation: stratified 5-fold cross-validation

The complete dataset (n = 318) was used in stratified 5-fold cross-validation, ensuring equal class distribution within each fold. Each fold was trained for 100 epochs, with the model achieving the highest validation accuracy selected. The best epoch and corresponding performance metrics—accuracy, precision, recall, F1-score, and AUC—are summarized in Table 3.

The model demonstrated stable performance across most folds, with 4 out of 5 achieving accuracy above 0.90 and AUC values above 0.93. One fold (Fold 5) showed relatively lower performance but still maintained an AUC of 0.8202. Detailed training curves for each fold are shown in Fig. 1, illustrating how both ac-

Table 3. Fold-wise metrics for stratified 5-fold cross-validation

Fold	Best epoch	Accuracy	Precision	Recall	F1-score	AUC
1	100	0.9375	0.9677	0.9091	0.9375	0.9423
2	17	0.8906	0.8462	0.9706	0.9041	0.9353
3	4	0.9219	0.9143	0.9412	0.9275	0.9618
4	33	0.9206	0.9375	0.9091	0.9231	0.9707
5	31	0.7460	0.7429	0.7879	0.7647	0.8202

All performance metrics were rounded to 4 decimal places.
AUC, area under the curve.

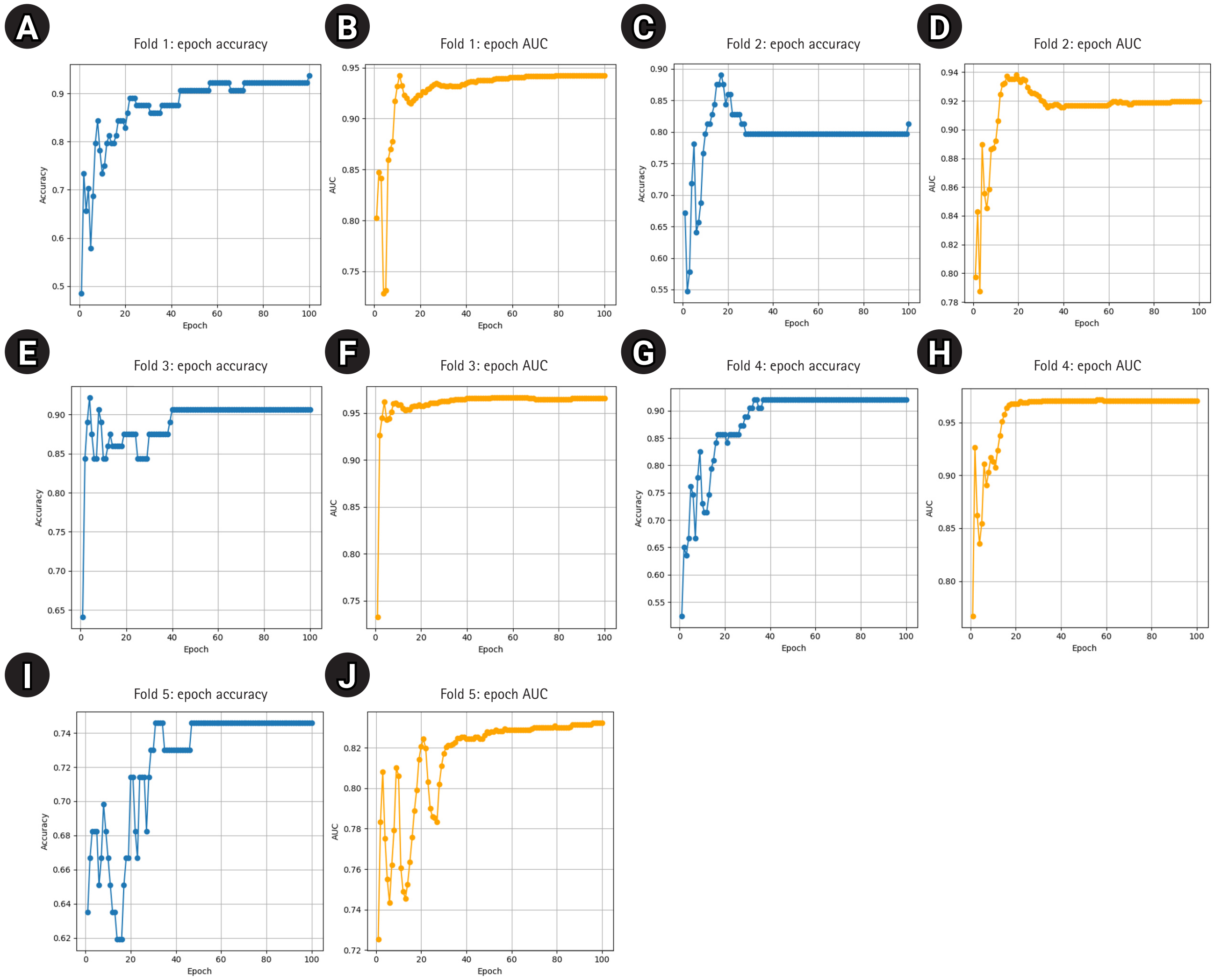


Fig. 1. (A–J) Epoch-wise accuracy and area under the receiver operating characteristic curve (AUC) for each fold.

curacy and AUC progressed and converged over the 100 epochs. The average epoch time per fold ranged from 8.3 to 8.6 seconds, confirming the computational efficiency of ResNet-18 in this small- to medium-scale medical imaging task.

External validation: hold-out test with ensemble voting

To assess generalizability, a separate hold-out test set of 20 images was created. The hold-out set included 10 class 0 and 10 class 1 images, selected via stratified sampling for balanced class representation. Detailed sample composition and individual model predictions are provided in Supplement 2. These samples were excluded from training and reserved solely for external validation. The remaining 298 images were used to train 5 ResNet-18 models using stratified 5-fold cross-validation. The best-performing model for each fold was selected based on the highest classification accuracy on the respective test fold. These models, trained on the reduced dataset, demonstrated consistently high performance: all folds achieved accuracy above 0.80 and AUC above 0.86, with 1 fold reaching 0.95 in both metrics (Table 4). These results indicate stable and effective training despite the reduced sample size. Detailed fold-wise metrics are presented below, and training curves are shown in Fig. 2.

Each of the 5 models was then used to independently predict the 20-image hold-out test set. Final ensemble predictions were made using soft voting, averaging the predicted probabilities for each class across the 5 models. The ensemble model correctly classified 18 out of 20 images, achieving an accuracy of 0.9000. Notably, all 10 class 1 images were correctly identified, resulting in perfect recall for early hair loss detection. Two class 0 images were misclassified as class 1. The ROC curve showed an AUC of 0.970, and the 95% CI for accuracy, calculated using the Wald method,

Table 4. Fold-wise metrics from training on the 298-image dataset

Fold	Best epoch	Accuracy	Precision	Recall	F1-score	AUC
1	11	0.8333	0.8621	0.8065	0.8333	0.8788
2	14	0.8000	0.7632	0.9063	0.8286	0.8650
3	23	0.9500	0.9143	1.0000	0.9552	0.9542
4	22	0.8475	0.8929	0.8065	0.8475	0.9182
5	23	0.8644	0.8966	0.8387	0.8667	0.9389

AUC, area under the curve.

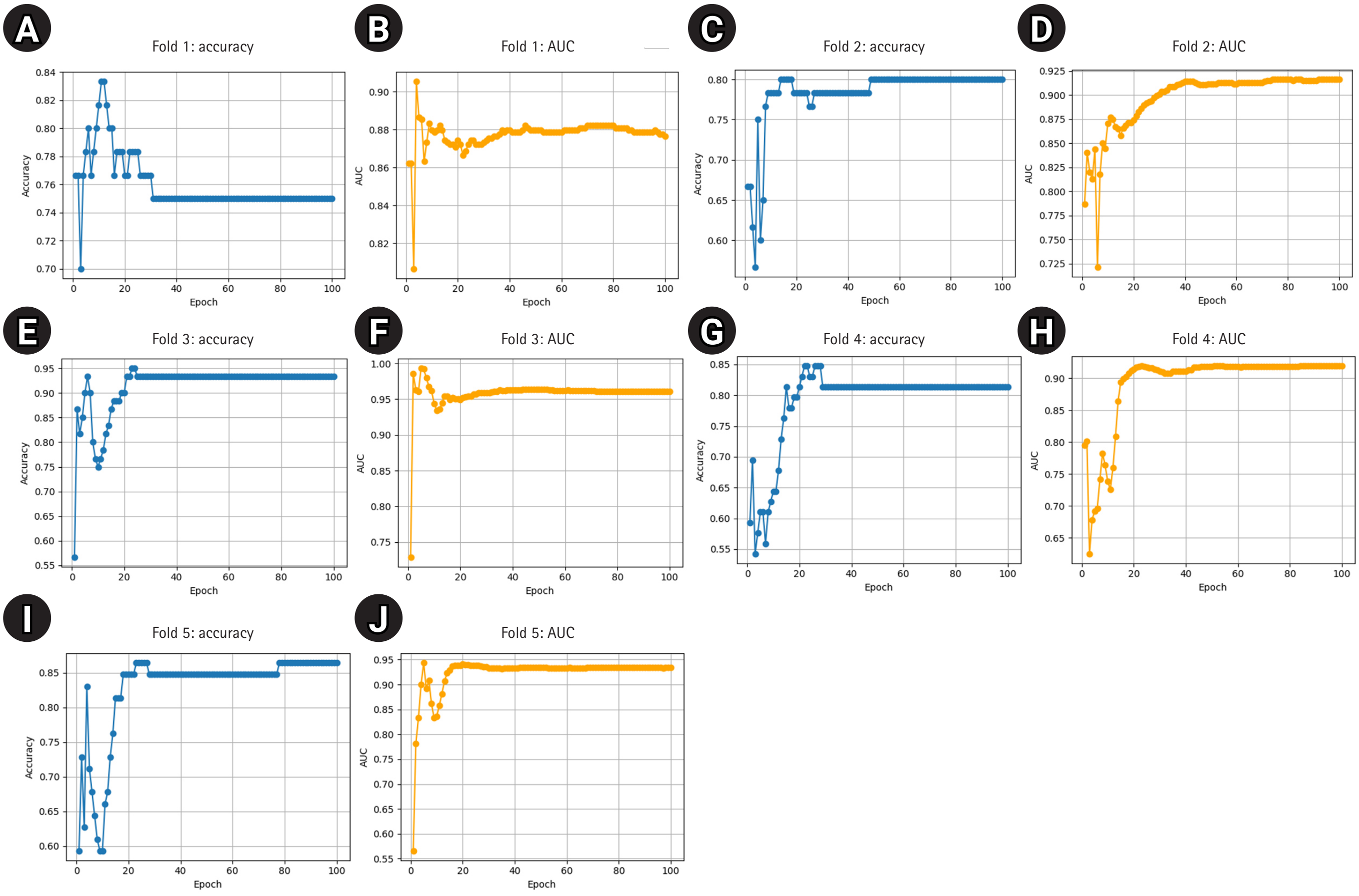


Fig. 2. (A–J) Epoch-wise accuracy and area under the receiver operating characteristic curve (AUC) for each fold trained on 298 images.

was 0.768 to 1.032.

The confusion matrix and ROC curve illustrating ensemble performance are presented in Fig. 3. Detailed predictions for each of the 20 hold-out images, including model-specific votes and ensemble results, are provided in Supplement 3.

Discussion

Key results

This study developed a ResNet-18-based deep learning model for classifying trichoscopic images by early-stage AGA severity using the BASP system. The model consistently demonstrated high performance in internal validation via stratified 5-fold cross-validation and maintained robust generalizability in external ensemble testing. High accuracy and AUC values in both validation settings confirmed the model's reliable discrimination between BASP 0 and BASP 1–3, even on previously unseen data.

Interpretation

The goal of this study was to assess the feasibility of deep learning for early-stage hair loss screening based on the BASP classification. The model showed consistently high performance across stratified internal folds, reflecting its ability to learn relevant patterns from trichoscopic images despite the limited data. Notably, 4 out of 5 internal folds achieved strong accuracy and AUC scores, with only 1 fold showing relatively reduced performance, likely due to incidental variation in class composition within that particular split. Nevertheless, this fold still maintained a respectable AUC of 0.8202, suggesting overall robustness across validation subsets.

This stability may be attributed to the binary simplification of BASP labels (BASP 0 vs. BASP 1–3), which reduced class fragmentation and enhanced the signal-to-noise ratio during training. Stratified 5-fold cross-validation further mitigated bias from class imbalance and ensured that every image contributed to both training and validation—a critical design choice for small datasets. Importantly, data augmentation was deliberately excluded, yet the model still exhibited stable convergence across folds (Figs. 1, 2), indicating that core patterns were sufficiently learnable from raw image features alone.

Performance on the external hold-out test set further validated the model's generalizability. Ensemble soft voting, based on 5 independently trained models, successfully classified 18 out of 20 images, achieving 90% accuracy and an AUC of 0.970. All early hair loss cases (class 1) were correctly identified, resulting in perfect recall. In clinical screening, such a low false-negative rate is crucial for timely intervention and minimizing missed diagnoses.

These findings collectively suggest that, with careful design, such as label restructuring, stratified sampling, and ensemble evaluation, even small, real-world clinical datasets can support the development of reliable deep learning models for early-stage disease detection. While the external test set was limited to 20 images, the ensemble approach helped compensate for this limitation by reducing model variance and strengthening prediction confidence.

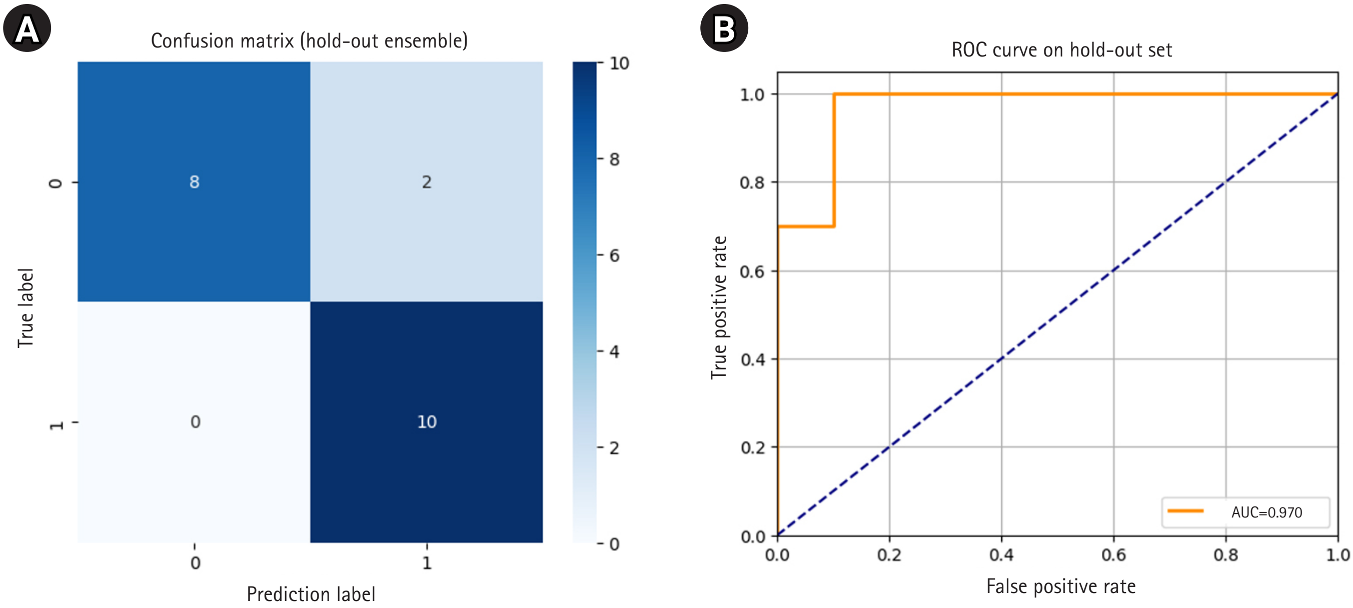


Fig. 3. Confusion matrix (A) and receiver operating characteristic (ROC) curve (B) for ensemble predictions on the hold-out set. AUC, area under the ROC curve.

Comparison with previous studies

While previous research has applied CNNs to dermatologic imaging, few studies have specifically targeted early-stage AGA or integrated BASP classification into model development. Most existing approaches rely on multiclass classification or segmentation tasks, which typically require larger datasets and extensive manual labeling.

In contrast, this study demonstrates that clinically meaningful classification can be achieved through strategic label simplification and robust validation strategies, even with limited data. By employing stratified cross-validation and ensemble-based external testing, this work addresses a common gap in medical artificial intelligence (AI) research: the reliance on internal metrics without assessment of generalizability to unseen data.

Limitations

The main limitation of this study is its small sample size (n = 318), which may restrict generalizability. In particular, images labeled as BASP 2 and 3 were underrepresented, potentially limiting the model's ability to learn nuanced patterns across progressive hair loss stages. A larger, more diverse dataset would support more robust training and enable a more comprehensive evaluation, including an expanded hold-out test set. All images were obtained from a single institution under specific imaging conditions, which may not capture the full variability seen in wider clinical settings, such as differences in scalp types, lighting, or trichoscopic equipment. Additionally, only one CNN architecture (ResNet-18) was evaluated. Comparative assessments across multiple architectures or training configurations could provide greater insight into model optimization.

Training hyperparameters, such as batch size and learning rate, were not systematically optimized. Although a batch size of 8 was initially tested, training did not converge effectively under that condition. A batch size of 128 was subsequently adopted based on successful convergence and was maintained throughout the study. Beyond this adjustment, no systematic exploration or substudy was conducted to identify optimal configurations for further improving model performance in this dataset.

Finally, the performance variability observed across cross-validation folds—most notably in Fold 5—highlights the model's sensitivity to class composition and sampling variation, underscoring the challenges of training on limited, imbalanced medical datasets.

Generalizability

Despite the limited dataset, ensemble-based hold-out testing demonstrated strong generalizability to unseen images. The model's high recall for early-stage AGA suggests potential clinical value as a screening tool, especially in resource-limited settings such as primary care or teledermatology. However, because all data were sourced from a single institution using one specific dermoscopy device, and all images were from Korean patients, broader generalization across different populations (varying in age, sex, ethnicity, and imaging equipment) will require future validation using multicenter, multi-ethnic, and multi-device datasets.

Suggestions for further studies

To build upon these findings and develop a practical diagnostic framework for early-stage AGA, several future directions are suggested.

First, multicenter data collection encompassing diverse populations, imaging devices, and clinical environments is essential to enhance generalizability and reduce bias related to demographics or equipment. Larger and more balanced datasets would also enable finer label granularity—for example, distinguishing BASP 1 (early) from BASP 2–3 (progressive)—to better reflect the clinical spectrum of hair loss.

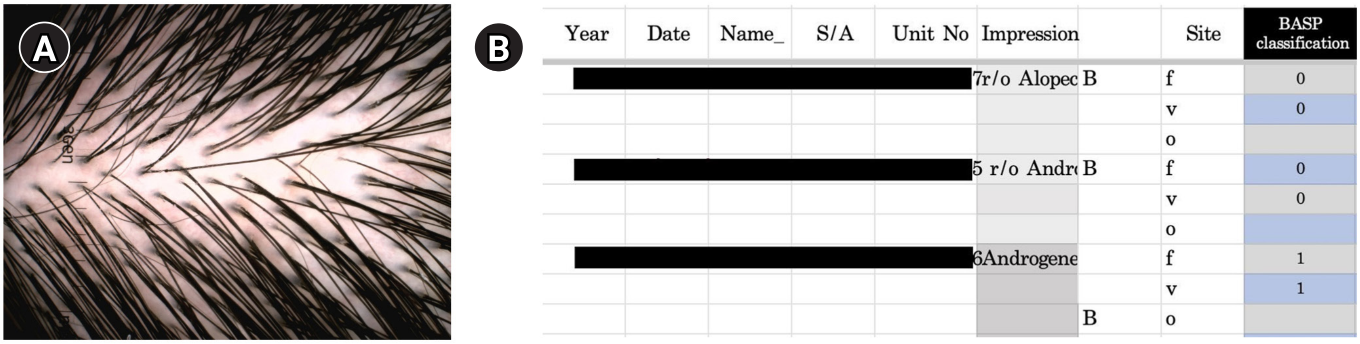
Second, comparative evaluation of alternative neural network architectures, such as EfficientNet, DenseNet, or vision transformers, should be performed to identify optimal trade-offs among diagnostic accuracy, computational efficiency, and deployment feasibility.

Third, integrating explainability techniques (such as Grad-CAM) and uncertainty quantification methods (like CIs or Monte Carlo dropout) may improve clinical trust and facilitate human–AI collaboration. Fairness metrics should also be monitored to assess potential bias across age, sex, or ethnicity subgroups.

Finally, real-world implementation studies in primary care or teledermatology—including workflow simulations and user feedback—will be vital for validating the model's practical utility and educational value in early diagnosis scenarios.

Conclusion

This study demonstrated the feasibility of a deep learning–based approach for early detection of AGA by leveraging BASP score classification. By simplifying the task to a binary distinction between non-hair loss (BASP 0) and early hair loss (BASP 1–3), the model achieved strong performance in both internal validation and ensemble-based external testing, without requiring data augmentation or extensive hyperparameter tuning. The use of stratified cross-validation and ensemble soft voting enabled robust learning even with a limited dataset, suggesting practical applicability in clinical screening scenarios. In particular, the high recall for early hair loss cases indicates strong potential for timely inter-



Year	Date	Name_	S/A	Unit No	Impression		Site	BASP classification
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	7r/o Alopec	B	f	0
							v	0
							o	
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	5 r/o Andr	B	f	0
							v	0
							o	
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	6Androgene		f	1
							v	1
						B	o	

Fig. 4. (A, B) Example of scalp image and Basic and Specific (BASP) classification data.

vention. These results support the integration of automated BASP-based assessment tools into dermatological workflows, promoting more standardized and objective evaluation of hair loss in clinical and teledermatology practice.

Furthermore, such deep learning–based systems may reduce the burden of manual labeling, minimize subjectivity in early hair loss diagnosis, and offer consistent alerts for possible AGA, thereby improving the accessibility, accuracy, and standardization of dermatologic care.

ORCID

Min Jung Suh: https://orcid.org/0009-0007-6087-7823
Sohyun Ahn: https://orcid.org/0000-0002-0116-3325
Ji Yeon Byun: https://orcid.org/0000-0003-4519-9474

Authors' contribution

Conceptualization: MJS, SHA, JYB. Data curation: MJS, JYB. Formal analysis: MJS. Funding acquisition: SHA, JYB. Methodology: MJS. Project administration: SHA. Visualization: MJS. Investigation: MJS. Resources: JYB, SHA. Software: MJS. Supervision: SHA, JYB. Writing–original draft: MJS. Writing–review & editing: MJS, SHA, JYB.

Conflict of interest

So Hyun Ahn has been an assistant editor of the journal since 2024, and Ji Yeon Byun has been an assistant editor of the journal since 2016. However, they were not involved in the editorial or peer-review process of this manuscript. Otherwise, no potential conflict of interest relevant to this article was reported.

Funding

None.

Data availability

Due to privacy concerns, the raw trichoscopic image data cannot be publicly shared (see Fig. 4 for a representative example).

Acknowledgments

None.

Supplementary materials

Supplementary files are available from https://doi.org/10.7910/DVN/BSXO6A

Supplement 1. Python-based training and evaluation code (Jupyter Notebook, .ipynb format).

Supplement 2. Hold-out set composition.

Supplement 3. Model-wise predictions and ensemble results for each image in the hold-out test set.

References

1. Starace M, Orlando G, Alessandrini A, Piraccini BM. Female androgenetic alopecia: an update on diagnosis and management. Am J Clin Dermatol 2020;21:69-84. https://doi.org/10.1007/s40257-019-00479-x
2. Lee JY, Kim CH, Lee WS. Relationship between illness behavior and hair loss pattern according to the basic and specific (BASP) classification. Ann Dermatol 2023;35:318-320. https://doi.org/10.5021/ad.21.085
3. Lee S, Lee JW, Choe SJ, Yang S, Koh SB, Ahn YS, Lee WS. Clinically applicable deep learning framework for measurement of the extent of hair loss in patients with alopecia areata. JAMA Dermatol 2020;156:1018-1020. https://doi.org/10.1001/jamadermatol.2020.2188
4. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115-118. https://doi.org/10.1038/nature21056
5. Kuhn M, Johnson K. Applied predictive modeling. Springer; 2013. https://doi.org/10.1007/978-1-4614-6849-3

6. Goodfellow I, Bengio Y, Courville A. Deep learning. MIT Press; 2016.

7. Ismail Fawaz H, Forestier G, Weber J, Idoumghar L, Muller PA. Deep learning for time series classification: a review. Data Min Knowl Discov 2019;33:917-963. https://doi.org/10.1007/s10618-019-00619-1

eISSN: 2234-2591
Open Access

Original article

Ewha Med J 2025;48(3):e45
https://doi.org/10.12771/emj.2025.00675

The Ewha Medical Journal

Machine learning for automated cause-of-death classification from 2021 to 2022 in Korea: development and validation of an ICD-10 prediction model

Seokmin Lee[*], **Gyeongmin Im**

Statistics Research Institute, Statistics Korea, Daejeon, Korea

Purpose: This study evaluated the feasibility and performance of a deep learning approach utilizing the Korean Medical BERT (KM-BERT) model for the automated classification of underlying causes of death within national mortality statistics. It aimed to assess predictive accuracy throughout the cause-of-death coding workflow and to identify limitations and opportunities for further artificial intelligence (AI) integration.

Methods: We performed a retrospective prediction study using 693,587 death certificates issued in Korea between January 2021 and December 2022. Free-text fields for immediate, antecedent, and contributory causes were concatenated and fine-tuned with KM-BERT. Three classification models were developed: (1) final underlying cause prediction (International Classification of Diseases, 10th Revision [ICD-10] code) from certificate inputs, (2) tentative underlying cause selection based on ICD-10 Volume 2 rules, and (3) classification of individual cause-of-death entries. Models were trained and validated using 2021 data (80% training, 20% validation) and evaluated on 2022 data. Performance metrics included overall accuracy, weighted F1 score, and macro F1 score.

Results: On 306,898 certificates from 2022, the final cause model achieved 62.65% accuracy (F1-weighted, 0.5940; F1-macro, 0.1503). The tentative cause model demonstrated 95.35% accuracy (F1-weighted, 0.9516; F1-macro, 0.4996). The individual entry model yielded 79.51% accuracy (F1-weighted, 0.7741; F1-macro, 0.9250). Error analysis indicated reduced reliability for rare diseases and for specific ICD chapters, which require supplementary administrative data.

Conclusion: Despite strong performance in mapping free-text inputs and selecting tentative underlying causes, there remains a need for improved data quality, administrative record integration, and model refinement. A systematic, long-term approach is essential for the broad adoption of AI in mortality statistics.

Keywords: Cause of death; Deep learning; International Classification of Diseases; Korean Standard Classification of Diseases-8th Revision

Introduction

Background

In 2023, the number of deaths in South Korea reached 352,511, representing a 32.4% increase since 2013. As one of the fastest-aging societies, South Korea is projected to experience continued surges in mortality, with more than 500,000 deaths expected by 2038, according to 2022 population projections [1]. These trends present major challenges for compiling cause-of-death statistics, which are vital for public health and demographic policy and necessitate improvements in data collection and analysis workflows.

Statistics Korea compiles these statistics based on physician-issued medical death certificates and family-reported death registrations. Causes are recorded in accordance with World Health Organization (WHO) guidelines, including the immediate cause, up to 3 antecedent causes, and other contributory physical conditions. Statistical reporting focuses on the underlying cause. To maintain international comparability, WHO specifies rules for underlying cause selection in the International Classification of Diseases, 10th Revision (ICD-10) Volume 2 [2].

However, physicians often lack sufficient information when completing death certificates. To address this, Statistics Korea

[*]Corresponding email: leon32@korea.kr
Received: June 18, 2025 Revised: July 23, 2025 Accepted: July 23, 2025

This study is an abridged version of the research project "Expanding the application of automated AI-based statistical classification," conducted by the Statistics Research Institute of Statistics Korea, available from: https://www.prism.go.kr/homepage/asmt/popup/1240000-202300033

supplements its data with 22 types of administrative records, such as national health insurance data, cancer registries, police investigations, autopsy reports, and infectious disease notifications. Cause-of-death coders, certified both internationally and domestically, review these materials to finalize the underlying cause. As mortality rises, the workload for these professionals also increases, increasing the risk of errors.

Meanwhile, artificial intelligence (AI) has begun to supplant human judgment in various sectors. Its potential to alleviate the growing burden of mortality statistics makes it a promising tool. This study explores the feasibility of leveraging AI for the compilation of cause-of-death data.

Objectives

This study investigates the ways in which AI can support the compilation of cause-of-death statistics and identifies its current limitations. Maintaining statistical quality amid rising mortality presents a complex challenge; thus, early assessment of AI's effectiveness is crucial. By analyzing real-world applications, we assess the potential of AI and propose strategies for its integration into future workflows.

Methods

Ethics statement

This study involved analysis of publicly available data; therefore, approval from an institutional review board and informed consent were not required.

Study design

This prediction study was conducted using publicly available data from Statistics Korea. The study design adheres to the TRIPOD-AI reporting guidelines for deep learning applications in medical research (development or prediction), available at https://www.tripod-statement.org/.

Setting

The study setting was the nationwide death registration system of the Republic of Korea. Data were obtained from national death certificates issued between January 1, 2021, and December 31, 2022, totaling 319,198 certificates in 2021 and 374,389 in 2022. The dataset thus represents a general population setting, covering deaths from all regions of Korea (urban and rural) and all levels of care, including hospital deaths with detailed medical information and home deaths that may have less detail.

Participants

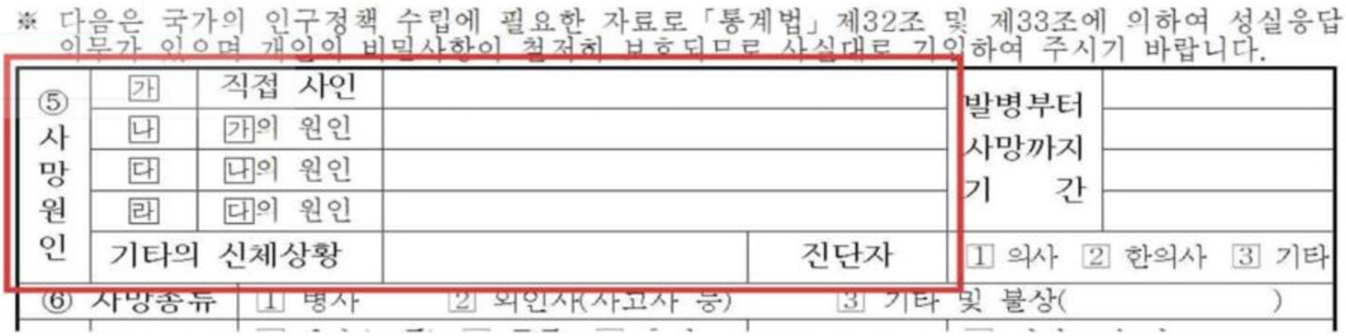
We effectively included the entire population of deaths in Korea during 2021–2022 for which a cause-of-death code was assigned.

Data source

Raw data were extracted from the Causes of Death Statistics database maintained by Statistics Korea. The study utilized 693,587 death certificate records (319,198 from 2021 and 374,389 from 2022), as coded by Statistics Korea, along with text input data from the underlying cause-of-death selection system (693,195 records) submitted by physicians or officers. Only records with matching case numbers across both datasets were selected and used as training data.

The cause-of-death section on Korean death certificates typically includes multiple entries: an immediate cause, intermediate causes, the underlying cause, and other contributing conditions (covering Part I and Part II of the WHO death certificate format) (Fig. 1).

However, coding rules require assignment of a single underlying cause-of-death code based on all this information. For modeling purposes, we concatenated the text from all cause-of-death fields to provide the model with full context. Specifically, we constructed a narrative summary by joining the entries in order—from immediate cause down to underlying cause, along with any other significant conditions—separated by punctuation. This ap-



※ 다음은 국가의 인구정책 수립에 필요한 자료로 「통계법」 제32조 및 제33조에 의하여 성실응답 의무가 있으며 개인의 비밀사항이 철저히 보호되므로 사실대로 기입하여 주시기 바랍니다.

⑤ 사망원인	가	직접 사인		발병부터 사망까지 기간	
	나	가의 원인			
	다	나의 원인			
	라	다의 원인			
	기타의 신체상황			진단자	1 의사 2 한의사 3 기타
⑥ 사망종류	1 병사	2 외인사(사고사 등)		3 기타 및 불상( )	

Fig. 1. Cause-of-death section on Korean death certificates.

proach mirrors how human coders review the entire certificate to determine the underlying cause.

Data preprocessing

Detailed preprocessing steps are presented in Supplement 1. Information on mortality statistics data—including inputs (cause of death), preprocessing, and outputs (automated candidate underlying causes)—is summarized in a diagram (Fig. 2)

Outcome variables/predictors

The outcome of interest was the underlying cause-of-death code, as defined by ICD-10, for each record. Thus, the model predicts a single label (the ICD-10 code) for each death certificate. The predictor variable is the textual content of the death certificate.

Bias

There was no bias in data collection or analysis, as all target data were included.

Study size

The study included the entire population of the Republic of Korea. No sample size estimation was necessary. Records with missing data were excluded from the analysis. Any death certificate lacking cause text was removed from the dataset before analysis, ensuring that the models did not encounter missing inputs.

Machine learning models

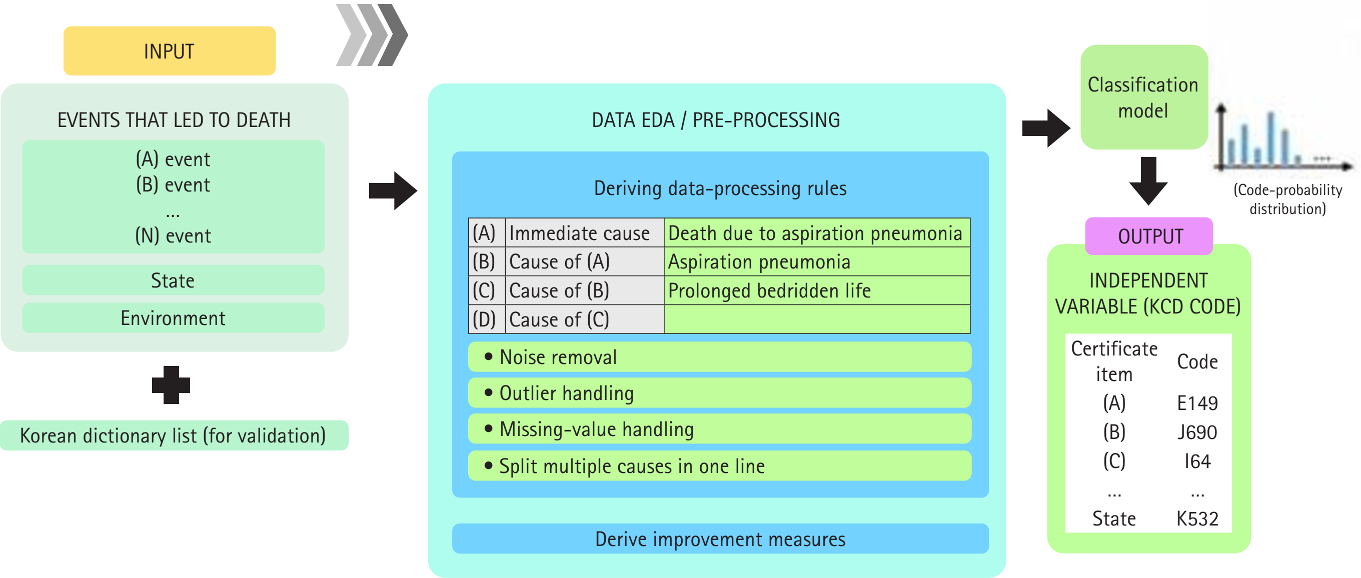
The Korean Medical BERT (KM-BERT) language model [3], a medical-domain pretrained BERT for Korean, was employed for natural language processing. KM-BERT is trained on over 60 million sentences and 1.16 billion Korean medical terms. Fine-tuning was conducted to achieve robust performance in medical term classification. Four models were trained and evaluated as described below:

Experiment 1: Development of a predictive model for the final underlying cause of death (8th Korean Standard Classification of Diseases and Causes of Death [KCD-8] code) using death certificate input fields

- Data source: All death certificate records with matching case numbers (N = 693,194).
- Model task: Directly predict the final underlying cause (KCD-8 code) from the input items on each death certificate.
- Objective: Quantify the rate at which forensic examiners revise the model's predicted classification for each cause-of-death category.

Experiment 2: Development of a baseline model for evaluating classification quality and similarity analysis of the automated underlying cause system

- Data source: A curated dataset of 310,034 death certificate records in which the automated underlying cause and the final underlying cause coincide.


INPUT
EVENTS THAT LED TO DEATH
(A) event
(B) event
...
(N) event
State
Environment
Korean dictionary list (for validation)
DATA EDA / PRE-PROCESSING
Deriving data-processing rules
• Noise removal
• Outlier handling
• Missing-value handling
• Split multiple causes in one line
Derive improvement measures
Classification model
(Code-probability distribution)
OUTPUT
INDEPENDENT VARIABLE (KCD CODE)

(A)	Immediate cause	Death due to aspiration pneumonia
(B)	Cause of (A)	Aspiration pneumonia
(C)	Cause of (B)	Prolonged bedridden life
(D)	Cause of (C)	

Certificate item	Code
(A)	E149
(B)	J690
(C)	I64
...	...
State	K532

Fig. 2. Information on mortality statistics data: inputs (cause of death), preprocessing, and outputs (automated candidate underlying causes).

- Model task: Using this curated dataset, a baseline classifier was trained to replicate the outputs produced by the automated underlying cause system.
- Objective: To validate the classification quality and analyze similarity between the model's predictions and the existing automated underlying cause results (with a particular focus on disease categories).

Experiment 3: Development of a knowledge-based classification model using only established resources

- Data source: Unstructured text fields from death certificates.
- Training resources: (1) An existing Korean language dictionary containing 74,123 entries; (2) The KCD-8 (based on ICD-10) master file, which has 89,584 Korean and English terminology entries (Supplement 2).
- Model task: Predict the classification code for each cause of death input field using only these curated vocabularies and domain knowledge.
- Objective: Evaluate the expected performance of a system built exclusively on pre-existing linguistic resources and domain expertise, without additional training examples.

Evaluation metrics

Accuracy, F1-macro, F1-weight, and loss were measured for each model.

Accuracy

The proportion of total predictions that the model classifies correctly. It is calculated as accuracy = [number of correct predictions]/[total number of predictions].

F1 macro

The arithmetic mean of the F1 scores computed independently for each class, giving equal weight to all classes regardless of their frequency. This metric reflects the model's balanced performance across both common and rare classes.

F1 weighted

The weighted average of per-class F1 scores, where each class's F1 is weighted by its support (the number of actual instances). This metric emphasizes performance on more prevalent classes, while still accounting for minority classes.

Statistical methods

Aside from the machine learning evaluation metrics described above, traditional statistical hypothesis testing for effect sizes or risk factors was not employed. The primary statistic of interest was model accuracy.

Results

Final underlying cause prediction (Experiment 1)

To determine the degree to which AI predictions of the underlying cause of death align with the actual causes, we trained the model solely on death certificates. We compared the underlying cause predicted by AI with the official underlying cause of death, which is determined using both death certificates and multiple administrative data sources. Of the 317,356 death certificates issued in 2021, a total of 140,080 records were used for training and validation. We then evaluated and compared results using 306,898 cases reported in 2022. The optimal classification model was reached at 300 steps (approximately 50 epochs). Upon evaluation of the 306,898 records from 2022, the model achieved a consistency rate of 62.65%. Approximately 37.35% of cases required modification, primarily due to the influence of external causes of death (Table 1).

Tentative underlying cause prediction (Experiment 2)

To evaluate the intermediate steps in generating cause-of-death statistics, we compared the tentative underlying cause—derived without administrative data—to the official results. Statistics Korea uses a term mapping table to automatically convert conditions written in Korean on death certificates into codes, then selects a tentative underlying cause according to WHO ICD-10 Vol. 2 guidelines. Measuring AI's ability to replicate this process may help reduce the burden of mapping table management.

A total of 310,034 refined datasets, in which the results before and after administrative data supplementation were identical, were used for training. The baseline model was trained using this curated data, and KM-BERT was employed as the language model. Optimal learning was achieved at 270 steps (approximately 90 epochs). Evaluation demonstrated high accuracy at 95.35% (Table 1). However, detailed results by classification item revealed areas of weakness in AI prediction. Of the 1,026 codes evaluated, 291 codes (e.g., E272, I7110, C740, C109) were highly reliable (F1-score ≥0.9), while 410 codes (e.g., L892, A1830, R578, R609) were unreliable (F1-score = 0). Unreliable results were particularly prominent for codes in "P. Prenatal and postnatal conditions," "S. Injury and poisoning," "M. Musculoskeletal diseases," "H. Eye and ear diseases," "Q. Congenital malformations," and "O. Pregnancy, childbirth, and postpartum diseases," all of which require supplemental administrative data. These findings indicate that specific disease categories require further model revision (Fig. 3).

Table 1. Summary of the validation and evaluation protocols for Experiments 1–4

Experiment	Input variables (X)	Output variables (Y)	Training/validation data	Evaluation data (comparison)	Accuracy	F1 score (weighted)	F1 score (macro)
1	Cause of death input information (5 items): direct cause, primary antecedent cause, secondary antecedent cause, tertiary antecedent cause, physiological condition	Final underlying cause classification code	2021 published data: input and classification codes (training: validation = 8:2)	2022 published final underlying cause classification codes (entire dataset)	0.6265	0.594	0.1503
2	Cause of death input information (5 items): direct cause, primary antecedent cause, secondary antecedent cause, tertiary antecedent cause, physiological condition	Final underlying cause classification code	2021 published data (curated subset where automatic underlying cause = final underlying cause): input and classification codes	2022 published automatic underlying cause classification codes	0.9535	0.9516	0.4996
3	Individual cause per record (single item)	Individual cause classification code	Existing Korean terminology dictionary+ICD-10 database master file	Cause-of-death selection system (simulation) results (20% test split) classification codes	0.7951	0.7741	0.925

ICD-10, International Classification of Diseases, 10th Revision.

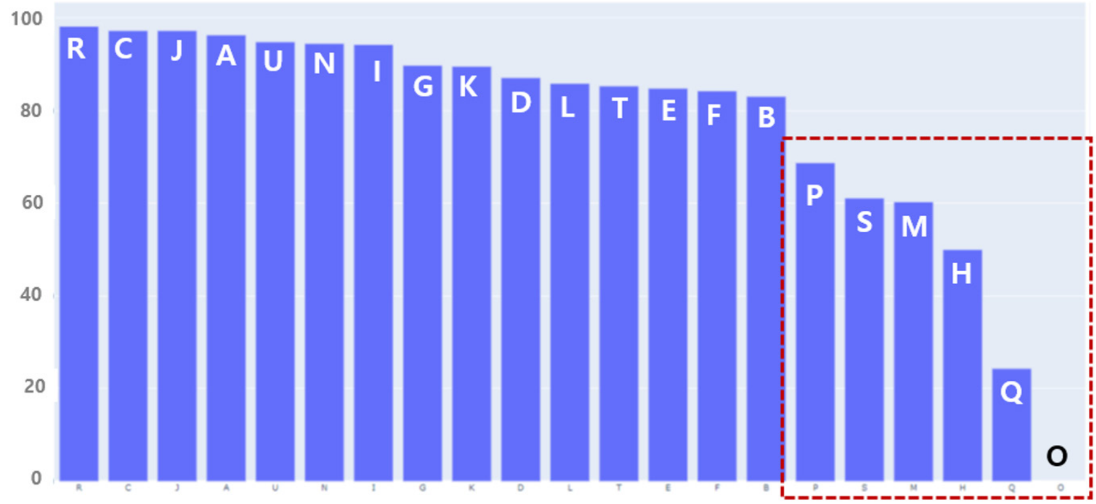


Fig. 3. Accuracy of artificial intelligence's prediction of the tentative underlying cause of death by major sections.

The number of misclassified causes of death was 18,116, with the majority being rare diseases such as anencephaly or cholera, which lacked sufficient training data. Many subcategories were also underrepresented, underscoring the need for expanded case data tailored to each classification purpose.

A typical error type occurred when frequently recorded causes of death, such as sepsis, were involved. In these cases, AI tended to predict categories for which it had more abundant training data. For example, when both sepsis and myelitis were listed, the model predicted M46.99 (unspecified inflammatory spondylopathy, unspecified site) instead of the more specific M46.59 (other infectious spondylopathy, unspecified site). This misclassification reflects the high frequency with which sepsis is recorded on death certificates for unspecified inflammatory spondylopathy.

Cause-of-death coding on death certificates (Experiment 3)

The accuracy of ICD-10 coding of causes of death described in natural language on death certificates was compared between artificial intelligence and term mapping table methods. Training data included the ICD-10 code list and the term mapping table used by Statistics Korea. The optimally fine-tuned classification model was learned at 1,550 steps (approximately 50 epochs). Evaluation showed that the AI coding accuracy was relatively low at 79.51%. However, the F1-Macro score was very high at 0.925, indicating strong agreement for items where prediction results were produced (Table 1).

Discussion

Key results

The process of producing cause-of-death statistics is primarily divided into 3 stages: (1) coding of natural language on death certificates, (2) selection of the tentative underlying cause of death according to ICD-10 Volume 2, and (3) modification of the final underlying cause of death based on administrative data. To determine what AI can predict at each step, we analyzed both the official statistical outcomes and AI prediction results. Experimentally, the model achieved 62.65% accuracy for the final underlying cause of death, 95.35% accuracy for the tentative underlying cause, and 79.51% coding accuracy for the cause of death [3].

Interpretation

Even for provisional causes of death where overall accuracy was high, unreliable results were observed for specific diseases. Understanding the overall volume of deaths is important, but so is managing trends for detailed causes and rare diseases. Thus, even rare diseases or deaths with low occurrence must be classified accurately. In this regard, our experiment highlights considerations relevant to the introduction of AI in this field.

AI misclassification cases

There are 3 major types of errors that AI can make when generating cause-of-death statistics. First, AI can confuse symptoms commonly listed on death certificates. For instance, sepsis often appears as a symptom secondary to various causes of death. When sepsis, spondylitis, and pneumonia are all present, the case should be classified as M46.59 (other infectious spinal diseases), yet it is frequently misclassified as M46.99 (unspecified inflammatory spinal diseases), a common error.

Second, errors may arise from insufficient understanding of causal relationships in disease coding. For example, diabetic renal failure should be classified as E14.28 (unspecified diabetes with renal complications), but the AI instead predicts E14.9 (unspecified diabetes without complications).

Third, errors occur when the AI lacks a nuanced understanding of natural language on death certificates. For instance, when the certificate states “동맥 관 개방,” it should be classified as Q25.0 (patent ductus arteriosus), but is instead classified as Q21.4 (aortopulmonary septal defect). Such errors often result from limited training data or ambiguity in language processing during AI calculation.

Limitations of applying AI to cause-of-death statistics

The classification for cause of death includes approximately 18,000 categories, far more detailed than the industrial classification system of about 1,200 categories. Accurate classification is crucial, not only for common causes but also for rare or infectious diseases, as these statistics are fundamental for public health policy. Therefore, a higher standard of accuracy is required compared to other statistical domains.

Considering these characteristics, applying AI to cause-of-death statistics poses significant challenges in predicting specific diseases or rare cases, since traditional modeling procedures may be inadequate. AI performance depends greatly on the diversity and volume of training data. Misclassification is particularly likely for rare or critical diseases with limited data.

To achieve successful AI application in this area, methods for correcting errors in specific cases must be considered. Supplementation with administrative data is also essential, and future AI models will require technology capable of verifying and integrating large-scale administrative records.

Directions of development for AI applications for cause-of-death statistics

Accurate prediction of specific diseases may benefit from a hybrid approach that combines case-based inference with established knowledge construction techniques, such as mapping tables or information retrieval methods. Additional experiments and advanced research are needed to build fine-tuned models using ensemble methods—leveraging existing machine learning algorithms capable of handling small sample sizes, large language models with extensive parameters, or specialized models focused on diseases and causes of death.

Furthermore, because cause-of-death statistics significantly impact public health policy, a long-term, systematic review of AI introduction is needed for each procedural step. Importantly, if the quality of death certificates—serving as the fundamental data source—does not improve, both the input quality and completeness of AI training data will suffer. It is therefore critical that death certificates are completed thoroughly and accurately.

Conclusion

The growing proportion of older adults in the population is causing a sharp rise in the number of deaths, which in turn may affect the quality of cause-of-death statistics. Statistics Korea utilizes an automated program to derive tentative causes of death from information listed on death certificates, followed by review and revision of the final cause based on administrative data. As the burden of statistical work increases, process quality may be compromised. Thus, integrating AI into the workflow represents a substantial opportunity for improvement.

However, because cause-of-death statistics involve more categories, higher complexity, and greater societal impact than other statistical fields, any changes in workflow must be accompanied by continuous, long-term review and analysis. In this study, we conducted a focused analysis using AI models trained on 2 years of data, confirming accuracy at each stage and identifying AI's limitations. This work is expected to lay the foundation for more effective AI applications in this area, ultimately supporting a healthier and safer society by enhancing the quality of cause-of-death statistics.

ORCID

Seokmin Lee: https://orcid.org/0000-0001-6642-2677
Gyeongmin Im: https://orcid.org/0009-0003-6655-6044

Authors' contributions

Conceptualization: SL, GI. Investigation: SL, GI. Methodology: SL, GI. Visualization: SL. Writing–original draft: SL. Writing–review & editing: SL, GI.

Conflict of interest

No potential conflict of interest relevant to this article was reported.

Funding

None.

Data availability

Research data and source codes can not be publicly shared due to the policy of Statistics Korea.

Acknowledgments

None.

Supplementary materials

Supplementary files are available from Harvard Dataverse: https://doi.org/10.7910/DVN/VT7IUV

Supplement 1. Data preprocessing and description of utilized data.

Supplement 2. KCD-8 DB master file.

References

1. Statistics Korea. Population prospects of Koreans and foreigners based on the 2022 population projections: 2022-2042 [Internet]. Statistics Korea; 2024 [cited 2025 Mar 11]. Available from: https://kostat.go.kr/board.es?mid=a20108080000&bid=11748&act=view&list_no=433289
2. World Health Organization. International Statistical Classification of Diseases and Related Health Problems, 10th Revision [Internet]. World Health Organization; 2019 [cited 2025 Mar 11]. Available from: https://icd.who.int/browse10/2019/en
3. Kim Y, Kim JH, Lee JM, Jang MJ, Yum YJ, Kim S, Shin U, Kim YM, Joo HJ, Song S. A pre-trained BERT for Korean medical natural language processing. Sci Rep 2022;12:13847. https://doi.org/10.1038/s41598-022-17806-8

eISSN: 2234-2591
Open Access

Health statistics

Ewha Med J 2025;48(3):e46
https://doi.org/10.12771/emj.2025.00689

The Ewha Medical Journal

Cause of death statistics in 2022 in the Republic of Korea

Jung-Hyun Oh*, Juhee Seo, Hyun Jung Park

Vital Statistics Division, Statistics Korea, Daejeon, Korea

Purpose: This study aimed to describe mortality trends in the Republic of Korea in 2022 by analyzing total deaths, crude and age-standardized mortality rates, as well as age- and sex-specific patterns and changes in cause-specific mortality. The analysis updates previous reports with newly available data from 2022.

Methods: A repeated cross-sectional analysis was performed using nationwide death certificate data collected through municipal administrative offices. Deaths occurring in 2022 were aggregated from reports filed over a 16-month period, spanning January 2022 to April 2023. Causes of death were classified according to the World Health Organization's International Classification of Diseases. Quality assurance was ensured through administrative record linkage across 22 databases and validation using an independent infant mortality survey. Descriptive statistics were employed to summarize the findings.

Results: In 2022, Korea recorded 372,939 deaths (the highest annual total since 1983), corresponding to a crude death rate of 727.6 per 100,000 population. This increase contributed to a net population decline of 123,751. Mortality rates rose across most age groups, with particularly marked increases among those aged 1–9 and those aged 80 or older. Coronavirus disease 2019 (COVID-19) became the third leading cause of death (31,280 deaths; 61.0 per 100,000), driven largely by the Omicron variant and heightened infection rates among older adults. Pancreatic cancer overtook stomach cancer in the mortality rankings. There were sharp increases in deaths attributed to Alzheimer's disease and diabetes. Although deaths from intentional self-harm declined, suicide remained a significant cause of death among younger individuals.

Conclusion: Korea experienced a record-high mortality rate in 2022, largely due to the impacts of COVID-19 and ongoing population aging. Notable shifts in cause-specific mortality were observed, including increases in deaths from Alzheimer's disease, diabetes, and pancreatic cancer, underscoring evolving public health challenges.

Keywords: Cause of death; COVID-19; Cross-sectional studies; Death certificates; Republic of Korea

Introduction

Background

In the Republic of Korea, mortality statistics are compiled in accordance with the Statistics Act and the Family Relations Registration Act. Death data, including certificates, are sourced from municipal administrative offices across the country. Cause-of-death statistics for 2014 [1], 2016 [2], 2018 [3], 2019 [4], 2020 [5], and 2021 [6] have previously been published. This report extends the analysis by incorporating newly available data from 2022.

Objectives

This study aimed to characterize mortality in Korea in 2022, including comprehensive assessment of total death counts, the crude death rate, the age-standardized death rate (ASDR), age- and sex-specific mortality rates, and recent trends in cause-specific mortality.

Methods

Ethics statement

Since this study utilized publicly accessible data, Institutional Review Board approval and informed consent were not required.

Study design

This study employed a nationwide, repeated cross-sectional design based on comprehensive death certificate data. Reporting

*Corresponding email: junghyunics@korea.kr

Received: June 20, 2025 Revised: July 20, 2025 Accepted: July 22, 2025

follows the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines, available at https://www.strobe-statement.org/.

Setting

Cause-of-death statistics in the Republic of Korea are compiled under the Statistics Act and the Act on the Registration of Family Relations. Data are collected using death notification forms and medical death certificates submitted to administrative offices nationwide. Deaths that occurred in 2022 were aggregated from reports filed over a 16-month period, spanning January 2022 through April 2023. All notifications and certificates were coded for the underlying cause of death according to the World Health Organization's International Classification of Diseases guidelines.

Participants (subjects)

Subjects included all death certificates for persons who died in Korea in 2022. While foreigners who died in Korea were documented separately, they were excluded from this analysis.

Data source/measurement

The data collection and analytical methods were consistent with those used in earlier cause-of-death studies, ranging from the 2016 report [2] through the 2018–2021 series [3-6]. For this analysis, all death certificates issued in 2022 for Korean residents of the Republic of Korea served as the primary data source [7].

To enhance the reliability of the mortality figures, 2 additional quality-assurance procedures were implemented. First, an independent infant mortality survey was conducted, since infant deaths are often underreported, by canvassing medical institutions and collecting cremation records from all crematoriums nationwide. Second, administrative record linkage was performed to validate the accuracy of recorded causes of death. This process involved cross-referencing each case against 22 distinct administrative databases, including national health insurance claims, the national cancer registry, police investigation files, autopsy reports, and other relevant records.

Variables

All causes of death were included as variables in the analysis.

Bias

No selection bias was present, as data from all subjects were included.

Study size

The entire population of the Republic of Korea was included; therefore, sample size estimation was unnecessary.

Statistical methods

Data were summarized using descriptive statistics only. No analytic statistical tests were performed.

Results

Number of deaths and crude mortality rate

In 2022, the total number of deaths in the Republic of Korea was 372,939, representing an increase of 55,259 deaths (17.4%) compared with 2021 (Fig. 1, Supplement 1). Male deaths numbered 196,465, which is an increase of 24,498 (14.2%) from the previous year, while female deaths totaled 176,474, up by 30,761 (21.1%). The average daily number of deaths rose to 1,022, which is 152 more than in the preceding year. The crude mortality rate (deaths per 100,000 population) was 727.6, reflecting an increase of 108.7 (17.6%) from 2021. Specifically, the male crude mortality rate reached 769.2 per 100,000 (an increase of 97.2, or 14.5%), while the female crude mortality rate rose to 686.2 per 100,000 (an increase of 120.2, or 21.2%). The male-to-female mortality ratio was 1.12, indicating that men's mortality rate was 1.12 times that of women. Both the total number of deaths and the crude mortality rate were the highest since national statistics were first compiled. The ASDR, which adjusts for age distribution, was 372.9, up by 55.3 from 2021 (Fig. 1, Supplement 1).

Deaths by sex and age group

Compared with 2021, the number of deaths increased most substantially among children aged 1–9 years (33.8%), adults aged 80 years and older (26.3%), those in their 70s (11.3%), and those in their 60s (10.2%) (Figs. 2, 3, Supplements 2, 3). Individuals aged 80 years and older accounted for 53.8% of all deaths in 2022, marking a 17.0 percentage-point increase from a decade earlier. Among male decedents, 40.7% were aged 80 or older—an increase of 16.7 percentage points over the past 10 years—while 68.3% of female decedents were in this age group, a 15.9 percentage-point rise. The male-to-female ratio of death counts peaked in the fifth and sixth decades of life, at 2.6. Age-specific mortality rates were lowest among children aged 1–9 years (11.3 per 100,000) and highest among those aged 80 years and older (9,237.2 per 100,000). The male crude mortality rate increased by 14.5% to 769.2 per 100,000, and the female rate rose by 21.2% to 686.2 per 100,000. Among males, age-specific mortality rates rose in the 1–9 and 10–19-year cohorts and in all age groups over 30. Among females, rates increased in the 1–9, 10–19, 20–29, and all age groups over 40. In every age stratum, males had higher

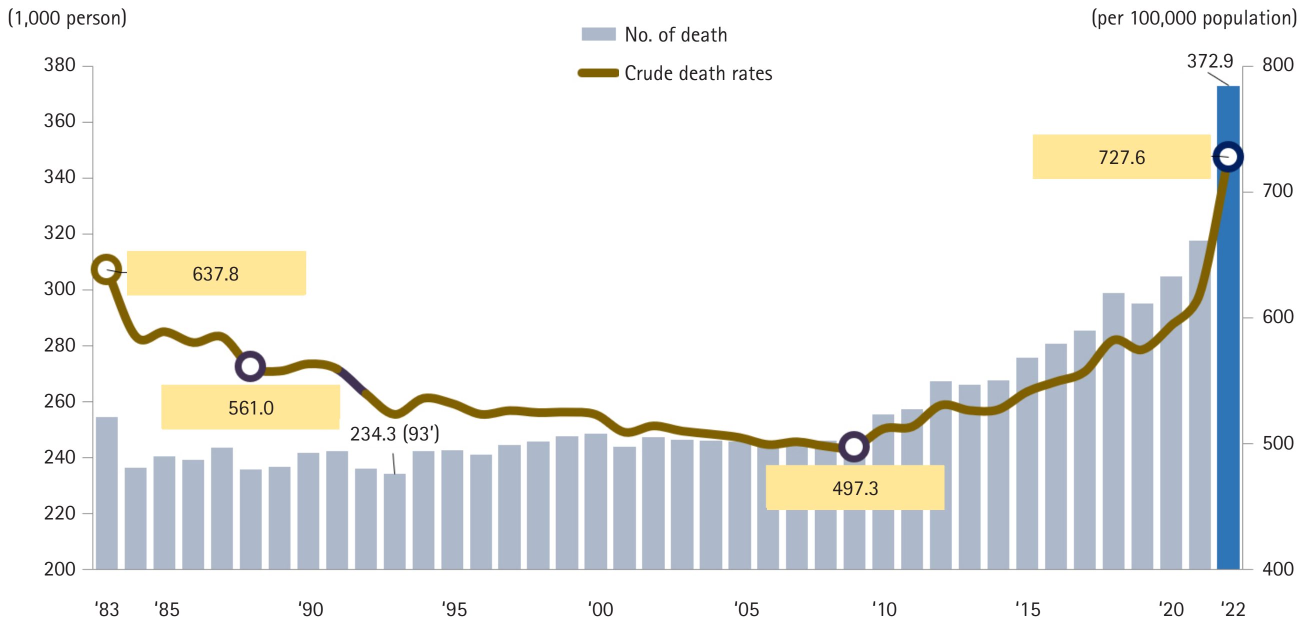


Fig. 1. The annual number of deaths and the crude death rate from 1983 to 2022 in Korea.

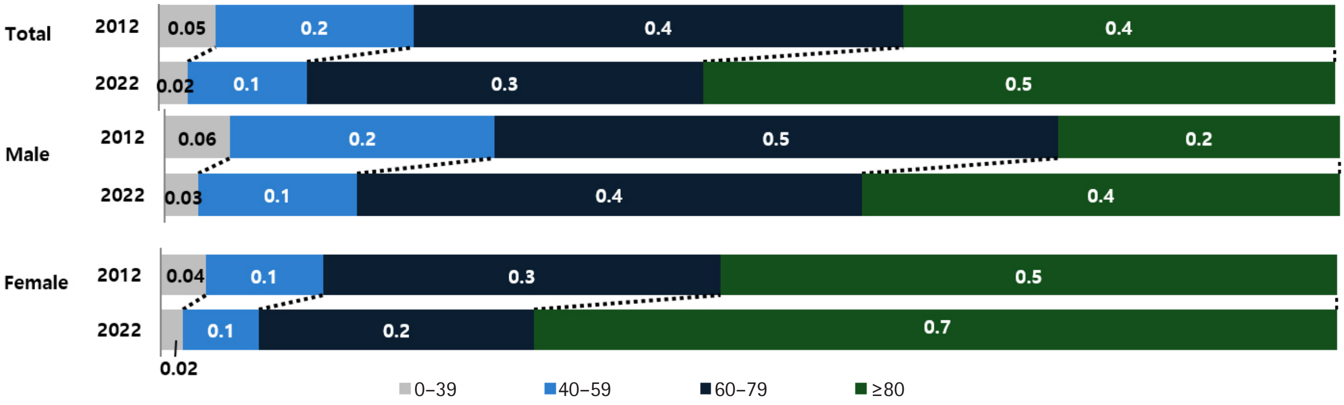


Fig. 2. Trends in sex- and age-specific proportions of deaths, 2012 vs. 2022 in Korea.

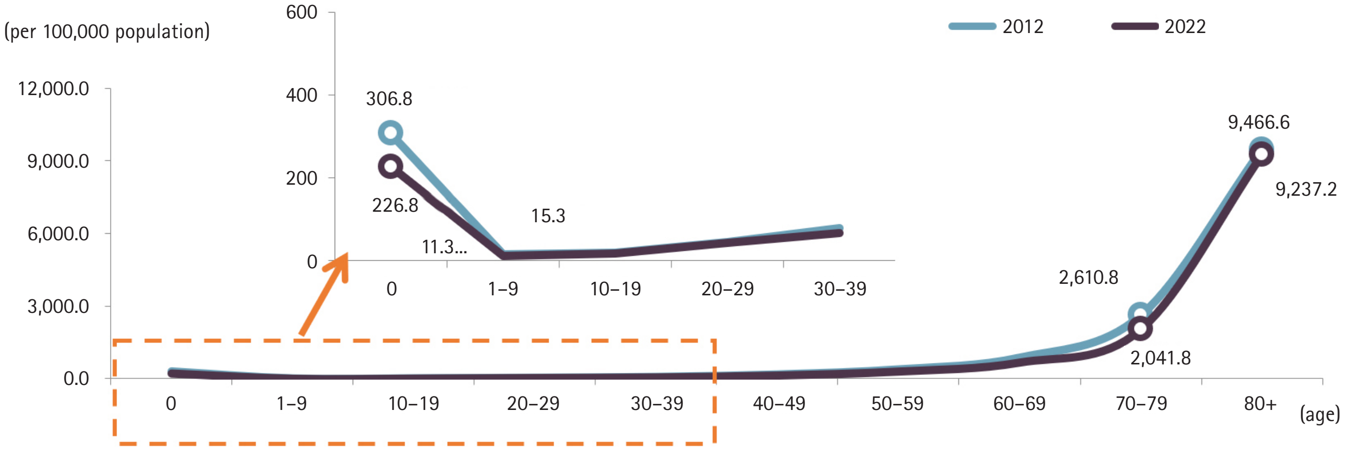


Fig. 3. Age-specific mortality rates by sex, 2012, 2021, and 2022 in Korea.

mortality rates than females, with the greatest disparity in the sixth decade, where the male rate was 2.7 times that of females.

Leading causes of death

In 2022, the top 10 causes of death were malignant neoplasms (cancer), heart disease, coronavirus disease 2019 (COVID-19), pneumonia, cerebrovascular disease, intentional self-harm, Alzheimer's disease, diabetes mellitus, hypertensive diseases, and liver disease (Fig. 4, Supplement 4). These 10 causes accounted for 67.4% of all deaths. The top 3—cancer, heart disease, and COVID-19—comprised 39.8% of total mortality, a 3.4 percentage-point decrease from 2021. COVID-19 appeared in the top 10 for the first time, ranking third. Malignant neoplasms and heart disease remained the leading causes by mortality rate, while the mortality rate for hypertensive diseases rose by 2.9 per 100,000 to 15.1 per 100,000. Compared to a decade earlier, pneumonia, Alzheimer's disease, and hypertensive diseases rose in ranking.

Among males, the 10 leading causes, in order, were malignant neoplasms, heart disease, COVID-19, pneumonia, cerebrovascular disease, intentional self-harm, diabetes mellitus, liver disease, chronic lower respiratory disease, and Alzheimer's disease. Male mortality rates exceeded those of females for cancer, pneumonia, intentional self-harm, diabetes mellitus, liver disease, and chronic lower respiratory disease. Among females, the top 10 were malignant neoplasms, heart disease, COVID-19, cerebrovascular disease, pneumonia, Alzheimer's disease, diabetes mellitus, hypertensive diseases, sepsis, and intentional self-harm. Females had higher mortality rates than males for heart disease, COVID-19, cerebrovascular disease, Alzheimer's disease, hypertensive diseases, and sepsis. For both sexes, malignant neoplasms were the leading cause, with the male cancer mortality rate 1.6 times that of females. COVID-19 and Alzheimer's disease appeared in the male top 10 for the first time, while the ranking of intentional self-harm fell for both sexes (males: 5th→6th; females: 7th→10th).

By age group, cancer was the leading cause of death among individuals aged 1–9 years and those aged 40 years and above; it was the second leading cause in the 10–19, 20–29, and 30–39 cohorts. Heart disease ranked second among those in their 60s and in all age groups except teenagers, consistently appearing in the top 5. COVID-19 ranked second for those aged 70 and older, third among children aged 1–9 years, and fourth in the 10–19 and 60–69 age groups. Pneumonia ranked fourth among those aged 80 and older and fifth in the 70s, highlighting increased risk in the oldest cohorts. Cerebrovascular disease was third in the 60s, fourth in the 70s, and fifth in the 40s, 50s, and those 80 and over. Liver disease ranked highest—third—among those in their 40s, and was fourth among those in their 30s and 50s. Intentional self-harm was the leading cause of death in teens through the 30s, second in the 40s and 50s, and fifth in the 60s.

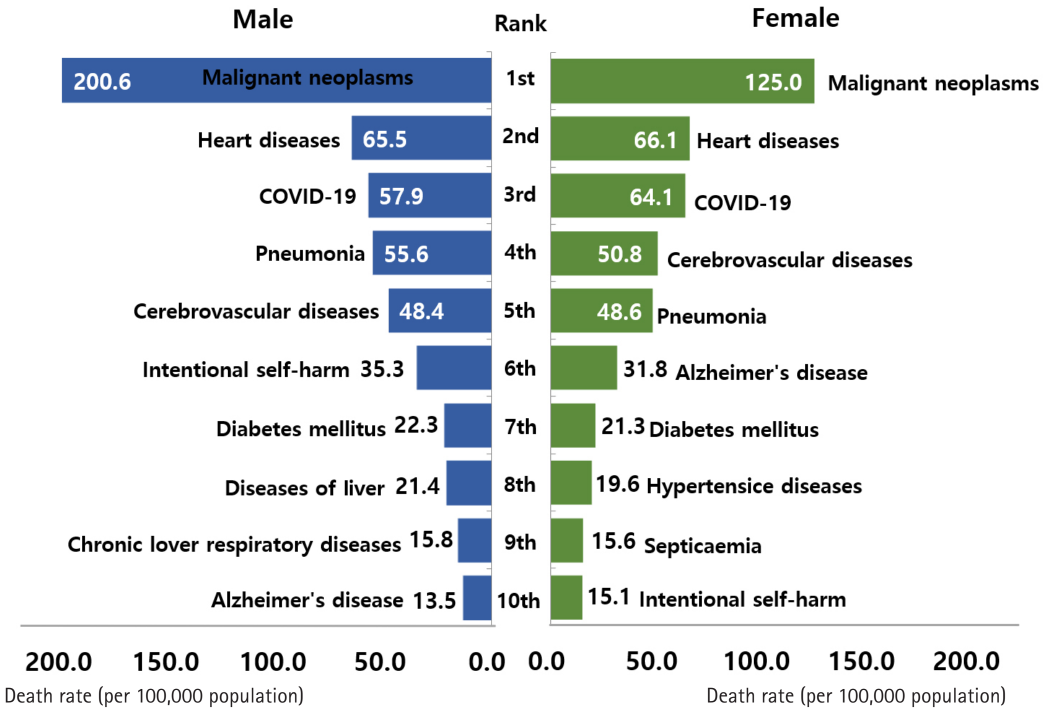


Fig. 4. The 10 leading causes of death by sex in 2021 in Korea.

Trends in cause-specific mortality rates

Overall trends

Compared with 2021, the most substantial relative increases in mortality rates (per 100,000 population) were observed for COVID-19 (522.8%), Alzheimer's disease (45.6%), diabetes mellitus (24.9%), hypertensive diseases (24.2%), pneumonia (17.3%), and cerebrovascular disease (12.6%) (Fig. 5, Supplement 5). In contrast, marked declines were noted for respiratory tuberculosis (–7.5%), transport accidents (–4.1%), and intentional self-harm (–3.2%). Over the past decade, the most pronounced long-term increases have occurred in Alzheimer's disease

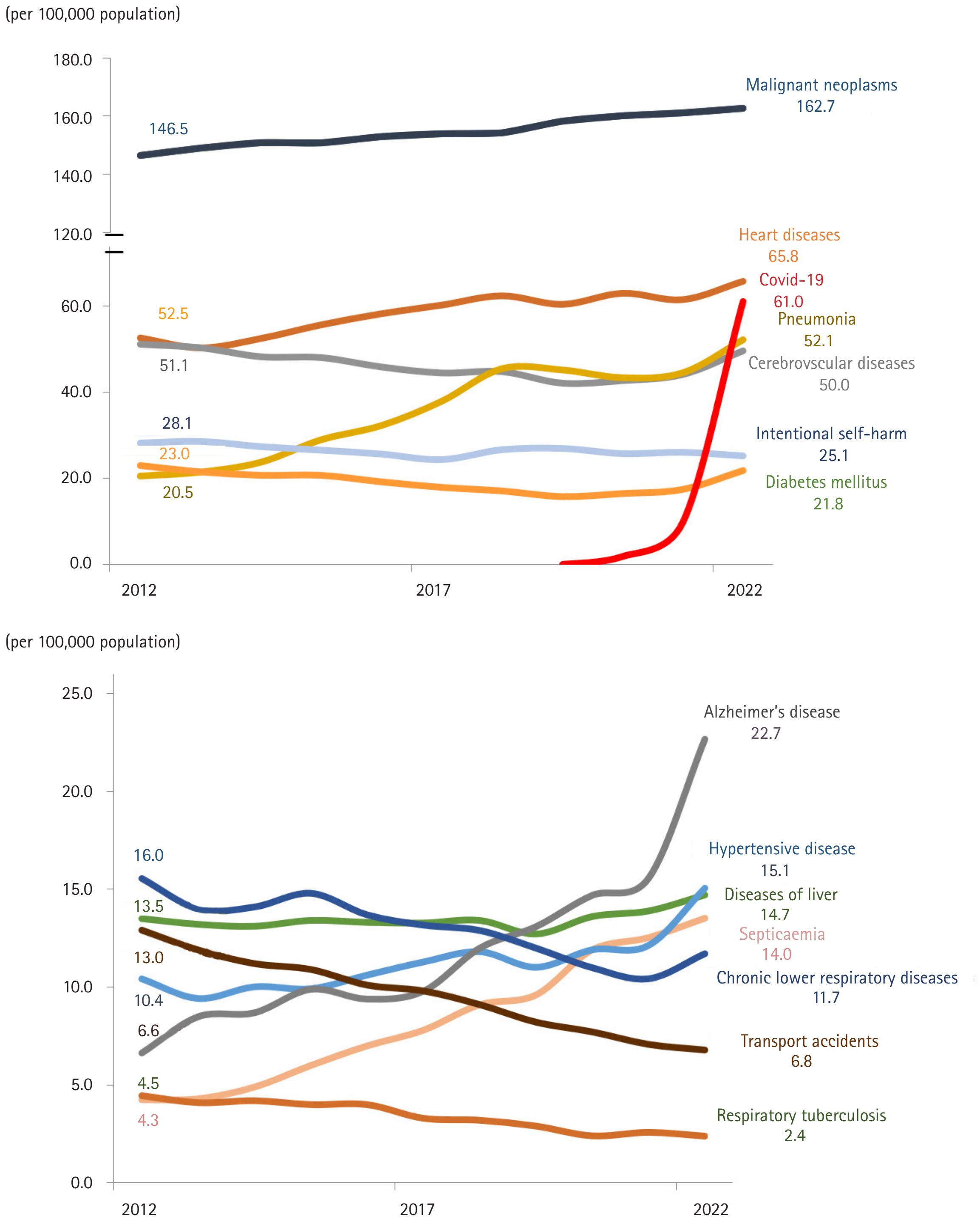


Fig. 5. Mortality rates trends for major causes of death in Korea.

(241.2%), sepsis (218.0%), pneumonia (154.4%), hypertensive diseases (44.7%), and heart disease (25.2%). Conversely, substantial long-term declines have been recorded for transport accidents (–47.6%), respiratory tuberculosis (–46.5%), chronic lower respiratory diseases (–24.7%), intentional self-harm (–10.5%), and diabetes mellitus (–5.0%).

Malignant neoplasms

The overall cancer mortality rate was 162.7 per 100,000, representing a 1.0% increase (1.6 per 100,000) from 2021. Lung cancer (36.3 per 100,000) remained the leading cause of cancer death, followed by liver cancer (19.9 per 100,000), colorectal cancer (17.9 per 100,000), pancreatic cancer (14.3 per 100,000), and stomach cancer (13.9 per 100,000) (Fig. 6, Supplement 6). Compared with the previous year, mortality rates increased for pancreatic cancer (5.8%), brain cancer (5.5%), and breast cancer (5.0%), but declined for uterine cancer (–4.3%), lung cancer (–1.5%), and stomach cancer (–1.3%). The cancer mortality rate for males (200.6 per 100,000) was 1.6 times higher than that for females (125.0 per 100,000). Among men, the highest mortality rates were recorded for lung cancer (53.7 per 100,000), liver cancer (29.1 per 100,000), and colorectal cancer (20.6 per 100,000); among women, the most fatal cancers were lung (18.9 per 100,000), colorectal (15.2 per 100,000), and pancreatic cancer (13.7 per 100,000). The most pronounced sex-specific disparity was seen in esophageal cancer, with a male-to-female ratio of 9.1, followed by lung cancer (2.8) and liver cancer (2.7). From 2021 to 2022, overall cancer mortality rates for both sexes increased by 1.6 per 100,000 (0.8% for men; 1.3% for women). Over the past decade, mortality rates have risen for pancreatic, lung, prostate, and breast cancers, while rates for stomach and liver cancers have declined. By age group, brain cancer was the leading cause among teenagers, leukemia in the 20s, stomach cancer in the 30s, breast cancer in the 40s, liver cancer in the 50s, and lung cancer among those aged 60 and above.

Circulatory system diseases

The mortality rate for circulatory system diseases was 134.7 per 100,000, with heart disease accounting for 65.8 per 100,000, cerebrovascular disease for 49.6 per 100,000, and hypertensive diseases for 15.1 per 100,000 (Fig. 7, Supplement 7). Compared with the previous year, rates increased for hypertensive diseases by 24.2%, cerebrovascular disease by 12.6%, and heart disease by 7.0%. Within heart disease, "other heart diseases" had the highest rate at 37.0 per 100,000. Female mortality from circulatory diseases (140.6 per 100,000) was 1.1 times higher than that of males (128.7 per 100,000). While women exhibited higher mortality for hypertensive and cerebrovascular diseases, men had a greater ischemic heart disease mortality rate (33.2 per 100,000 versus 24.3 per 100,000 for women). From 2021 to 2022, circulatory disease mortality increased by 12.6 per 100,000 (10.8%) among men and 13.8 per 100,000 (10.9%) among women. Mortality from circulatory diseases rose consistently with age, and in every age group, heart disease, cerebrovascular disease, and hypertensive diseases were the top 3 causes. Among individuals in their 40s through 60s, ischemic heart disease predominated, whereas in teenagers, those in their 20s, and those aged 70 and above, other heart diseases were most frequent.

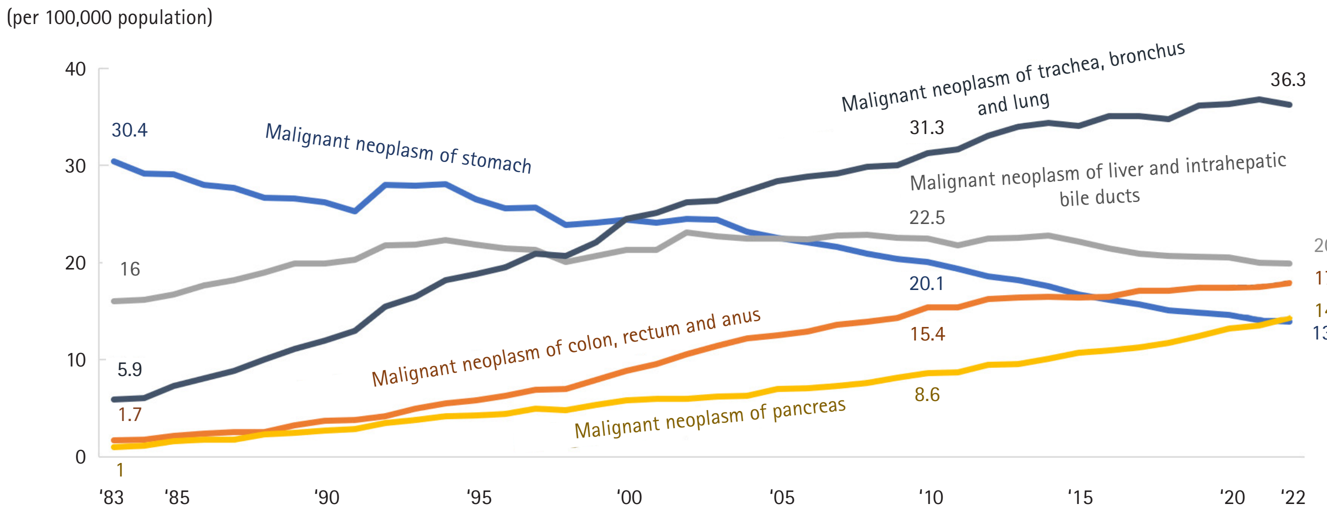


Fig. 6. Trends in mortality from malignant neoplasms by organ site from 1983 to 2022 in Korea.

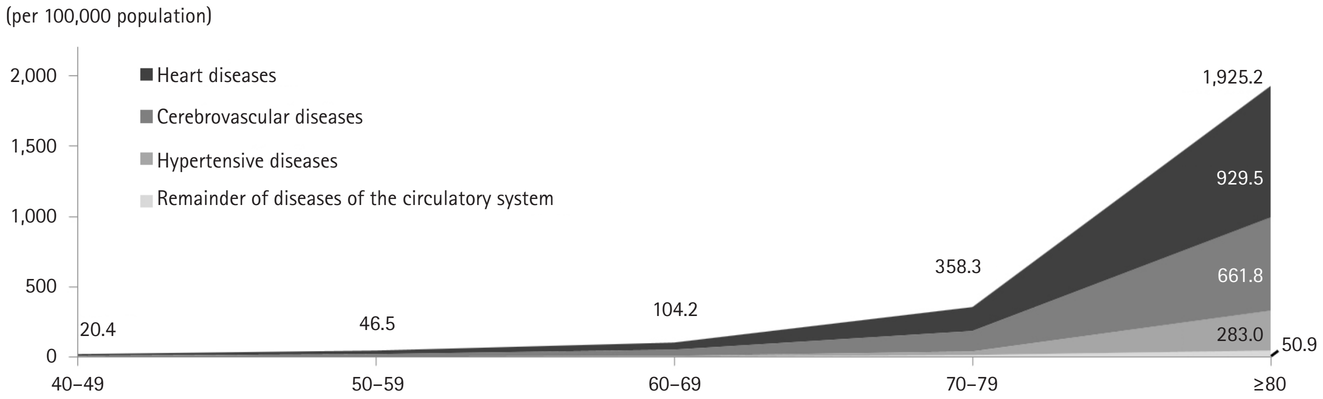


Fig. 7. The mortality rate due to circulatory system diseases by age in 2022 in Korea.

External causes of death (including accidents)

Non-disease external causes accounted for 7.2% of all deaths (26,688 cases), down 1.1 percentage points from 8.2% in 2021 (Figs. 8, 9, Supplements 8, 9). The mortality rate for external causes was 52.1 per 100,000, a 2.2% increase from the prior year. The leading external causes were suicide (25.2 per 100,000), transport accidents (6.8 per 100,000), and falls (5.3 per 100,000). Mortality rates declined for homicide (–10.0%), transport accidents (–4.1%), and suicide (–3.2%), but rose for fire-related accidents (9.2%), poisoning (4.7%), and drowning (2.5%). The external-cause mortality rate for males (71.4 per 100,000) was 2.2 times that for females (32.9 per 100,000). The highest male-to-female ratios were observed for drowning (3.2), transport accidents (3.0), and falls (2.6). By age, homicide (such as abandonment) (4.0 per 100,000), falls (1.2 per 100,000), and transport accidents (0.4 per 100,000) were most common among infants (0 years); in children aged 1–9 years, homicide (0.6) and equal rates (0.4) for transport, falls, and drowning accidents were seen; among individuals aged 10–79 years, suicide and transport accidents were most frequent; and in those aged 80 and older, suicide (60.6), falls (42.8), and transport accidents (29.3) per 100,000 were predominant.

Intentional self-harm trends

In 2022, there were 12,906 deaths by intentional self-harm, a decrease of 446 cases (3.3%) compared to 2021. Monthly declines were especially notable in March (–16.0%), June (–15.3%), and February (–13.1%). The average daily number of suicides was 35.4. The mortality rate for intentional self-harm was 25.2 per 100,000, a reduction of 0.8 (3.2%) from the previous year. Rates increased in the 40s (2.5%) and among teenagers (0.6%), but decreased in the 70s (–9.6%), 20s (–9.2%), 30s (–7.2%), 60s (–4.7%), 50s (–3.6%), and those aged 80 and above (–1.1%). Male intentional self-harm mortality (35.3 per 100,000) was 2.3 times higher than that for females (15.1 per 100,000). Both sexes experienced declines (males –1.7%; females –6.4%), with the lowest male-to-female ratio in the teenage group (1.1) and the highest in those aged 80 and older (3.8). Intentional self-harm remained the leading cause of death among those aged 10–39, and the second leading cause among those in their 40s and 50s.

Alcohol-related mortality

In 2022, there were 5,033 deaths attributable to alcohol, averaging 13.8 deaths per day and representing an increase of 105 deaths compared to 2021 (Fig. 10, Supplement 10). The alcohol-related mortality rate was 9.8 per 100,000, marking a 2.3% increase from the previous year. Male alcohol-related mortality rose in the 20s, 30s, 50s, and 70s, while female rates increased in all age groups except those in their 30s. The male mortality rate (16.7 per 100,000) was 5.7 times higher than that for females (3.0 per 100,000). Alcohol-related mortality rates increased steadily after the 30s, peaking in the 50s before declining.

Dementia-related mortality

Deaths attributed to dementia reached 14,136 in 2022, representing a 36.6% increase compared to 2021. The dementia mortality rate was 27.6 per 100,000, an increase of 7.4 (36.8%) (Fig. 11, Supplement 11). Female dementia mortality (38.0 per 100,000) was 2.2 times that of males (17.1 per 100,000). Both sexes experienced substantial year-over-year increases in dementia-related mortality (males 32.9%; females 38.5%).

COVID-19 mortality

In 2022, COVID-19 was responsible for 31,280 deaths, accounting for 8.4% of all fatalities. This number surpassed the 26,250 COVID-19 deaths recorded in 2021 (Fig. 12, Supplement

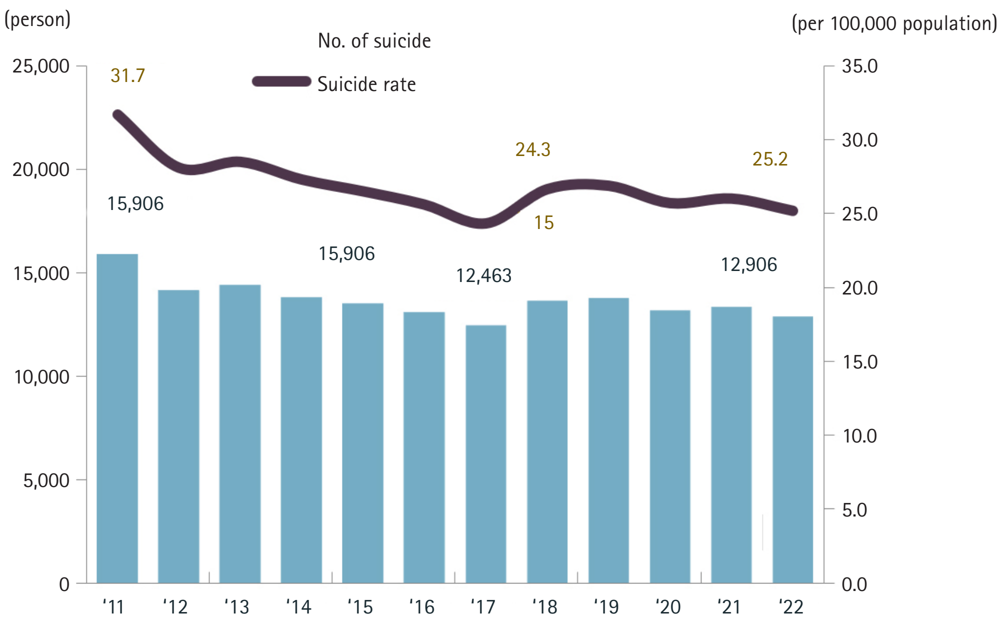


Fig. 8. Number of deaths and mortality rate due to intentional self-harm, 2011–2022 in Korea.

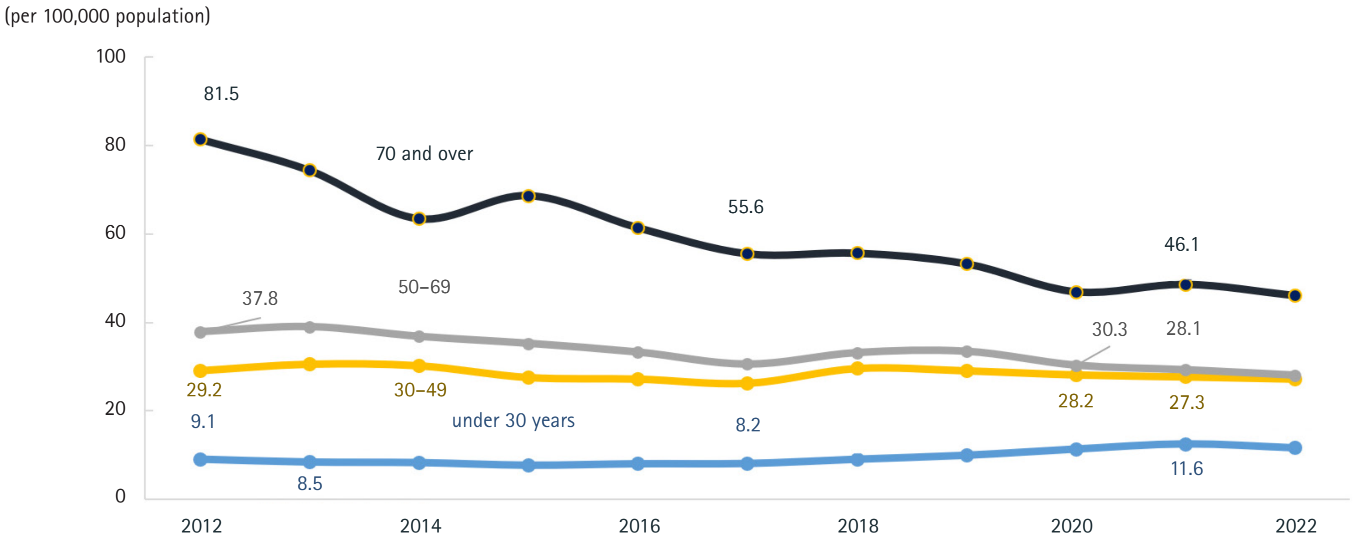


Fig. 9. Age-specific intentional self-harm rates, 2012–2022 in Korea.

12). The COVID-19 mortality rate reached 61.0 per 100,000, representing a 51.2-point (522.8%) increase over the previous year. Mortality rose sharply with age, peaking at 946.0 per 100,000 among individuals aged 80 years and older. Monthly COVID-19 deaths were highest in March (10,955), followed by April (6,875).

Discussion

The year 2022 marked a significant demographic shift for Korea, with the highest number of recorded deaths since 1983, totaling 372,937, an increase of 55,259 compared to the previous year. Simultaneously, the crude mortality rate reached an all-time high of 727.6 per 100,000 population (Fig. 1). This substantial rise in mortality, combined with a birthrate of 249,186 in 2022, resulted in a net population decrease of 123,751 [7]. Notably, this decline

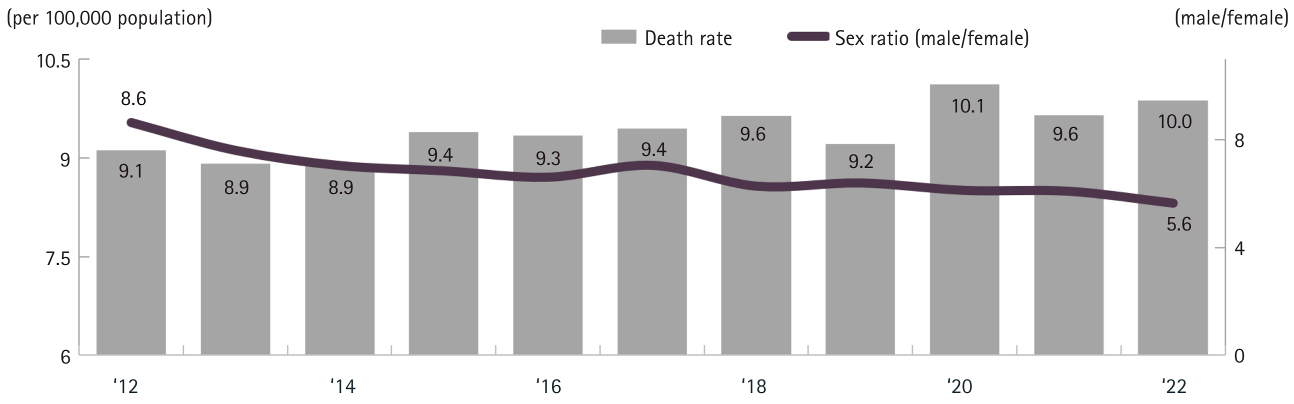


Fig. 10. Trends in the sex ratio of alcohol-related mortality, 2012–2022 in Korea.

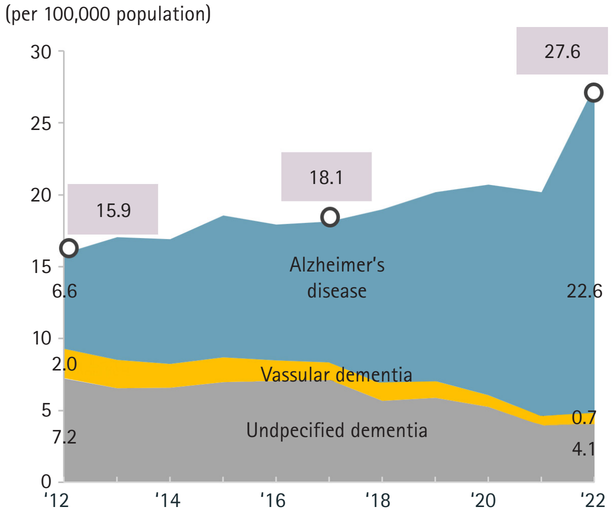


Fig. 11. Trends in mortality rates due to dementia by cause, 2012–2022 in Korea.

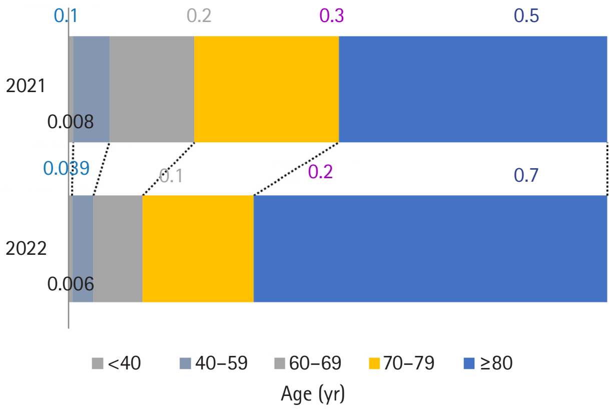


Fig. 12. COVID-19-related deaths per 100,000 population by age and sex, 2021 and 2022 in Korea.

was more than double the decrease observed in 2021 (57,080). Korea's total fertility rate in 2022 was just 0.78, the lowest in the world, suggesting that population decline will likely continue unless there is a dramatic increase in birth rates.

Age-specific mortality trends

Analysis of age-specific mortality rates per 100,000 population revealed a decrease of 6.0 deaths in the 0-year-old cohort and a reduction of 0.8 deaths among individuals in their 30s. In contrast, mortality rates rose in all other age groups. The most pronounced increase occurred in the 1–9 age group, with a rise of 42.0 deaths, while those aged 80 and above experienced an increase of 17.7 deaths (Fig. 3).

Leading causes of death

Among the causes of death, COVID-19-related fatalities rose sharply to 31,280 in 2022, a substantial jump from 5,030 in 2021. This increase resulted in a COVID-19 mortality rate of 61.0 per 100,000, accounting for 8.4% of all deaths and making it the third leading cause of death for both men and women, underscoring the rapid escalation of pandemic-related mortality (Fig. 4). The surge in COVID-19 deaths is primarily attributed to the high transmissibility of the Omicron variant and widespread community infection, particularly among the elderly.

The Omicron (BA.1) variant became the dominant strain in Korea in early 2022 [8], resulting in a dramatic surge in confirmed cases after January, with a peak of 9,959,368 confirmed cases in March. Notably, 55.4% (14,735/26,593) of all COVID-19 deaths in 2022 occurred in March and April—a period characterized by increased infections among older adults [9]. Although Korea achieved high vaccination rates—as of May 31, 2022, the proportions of the Korean population with complete vaccination and an additional booster shot were 86.8% and 66.9%, respectively, making Korea one of the countries with the highest vaccination rates worldwide [10]—a considerable number of deaths still occurred, particularly in highly vulnerable groups such as those over 80 years old, who experienced elevated fatality rates in breakthrough infections due to their relatively compromised immune systems [11]. During 2020–2021, Korea successfully suppressed community transmission through stringent containment measures such as social distancing. However, all social distancing restrictions were lifted in mid-April 2022 after a confirmed decline in the epidemic curve. It is important to note that this decision was made after the Omicron surge had peaked (mid-March), and that targeted prevention strategies for high-risk groups continued even after the lifting of restrictions.

Alzheimer's disease has shown a steady increase since entering the top 10 causes of death in 2018, rising from 15.6 to 22.7 deaths per 100,000 population by 2022 (Fig. 5).

Among the various causes of death, diabetes surpassed liver disease in the rankings for men. Traffic accidents and septicemia also fell out of the top 10 causes. For women, intentional self-harm declined in ranking relative to hypertensive diseases and septicemia, settling at the 10th position (Figs. 4, 8). The ongoing decrease in intentional self-harm among women compared to men remains difficult to explain and requires further observation to determine if this trend will persist.

Cancer mortality trends

Regarding cancer-related deaths, the rankings for lung cancer, liver cancer, and colorectal cancer remained unchanged from 2021. However, pancreatic cancer notably surpassed stomach cancer for the first time, rising to fourth place (Fig. 6). This pattern suggests a continuing decrease in stomach cancer deaths and a sustained increase in pancreatic cancer deaths in the future. While liver cancer deaths are expected to decline, colorectal cancer mortality is projected to remain stable. The decline in stomach cancer mortality can be attributed to the inclusion of gastroscopy in national health screening programs and increased screening participation, which have promoted earlier detection and shifts in dietary habits [12]. Similarly, the inclusion of occult blood tests in national screening and the widespread adoption of colonoscopy have helped prevent an increase in colorectal cancer mortality through early detection [13]. Lung cancer deaths have shown a slight decrease (Fig. 6), which may be attributable to biennial chest X-ray screening in national health check-ups. The application of artificial intelligence to chest X-ray interpretation is expected to further enable earlier lung cancer detection in the future [14].

Age as a determinant of mortality

Age is a primary determinant of mortality in South Korea. Older adults are especially vulnerable to chronic diseases such as diabetes and hypertension, as well as infectious diseases like septicemia and pneumonia. In this population, diminished immune function amplifies the burden of these illnesses. Notably, the prevalence of diabetes increases with age and is a major contributor to rising mortality among those aged 70 and above. The diabetes-related mortality rate per 100,000 population has climbed in recent years, from 16.5 in 2020 to 17.5 in 2021, and sharply to 21.8 in 2022 (Fig. 5).

Declining traffic accident fatalities and stable alcohol-related deaths

Traffic accident fatalities have shown a consistent decline (Fig. 5), a trend attributed to interventions such as the introduction of safety regulations in child protection zones in 1995, amendments to the Road Traffic Act in 2020, stricter penalties for drunk driving, and heightened public awareness of road safety. Alcohol-related deaths have remained relatively stable at 9.9 per 100,000 population, only a slight change from 9.6 in 2021 (Fig. 10). Without the implementation of robust national prohibition policies—such as banning alcohol consumption scenes in media, enforcing stricter license suspensions for drunk driving, and imposing harsher legal penalties for alcohol-related accidents—significant reductions in alcohol-related mortality are unlikely in the near term.

Conclusion

Korea experienced a marked increase in overall mortality in 2022, reaching unprecedented levels, mainly due to the emergence of COVID-19 as the third leading cause of death and the impact of an aging population. Notable shifts in mortality patterns included pancreatic cancer surpassing stomach cancer and significant increases in deaths related to Alzheimer's disease and diabetes.

ORCID

Jung-Hyun Oh: https://orcid.org/0009-0006-8272-5815
Juhee Seo: https://orcid.org/0000-0002-4426-5049
Hyun Jung Park: https://orcid.org/0009-0008-1417-5393

Authors' contributions

Conceptualization: JHO, JS, HJP. Investigation: JHO. Methodology: JHO, JS, HJP. Visualization: JHO. Writing–original draft: JHO. Writing–review & editing: JHO, JS, HJP.

Conflict of interest

No potential conflict of interest relevant to this article was reported.

Funding

None.

Data availability

Data file is available from Harvard Dataverse: https://doi.org/10.7910/DVN/DPXEXD

Dataset 1. Descriptive data on the cause of death in 2022 of Korean people, publicized by Statistics Korea.

Acknowledgments

None.

Supplementary materials

Supplementary materials are available from https://doi.org/10.12771/emj.2025.00689

Supplement 1. The annual number of deaths and the crude death rate in Korea from 1983 to 2022.

Supplement 2. Trends in sex- and age-specific proportions of deaths, 2012 vs. 2022 in Korea.

Supplement 3. Age-specific mortality rates by sex, 2012, 2021, and 2022 in Korea.

Supplement 4. The 10 leading causes of death by sex in 2021 in Korea.

Supplement 5. Mortality rates trends for major causes of death in Korea.

Supplement 6. Trends in mortality from malignant neoplasms by organ site from 1983 to 2022 in Korea.

Supplement 7. The mortality rate due to circulatory system diseases by age in 2022 in Korea.

Supplement 8. Number of deaths and mortality rate due to intentional self-harm, 2011–2022 in Korea.

Supplement 9. Age-specific intentional self-harm rates, 2012–2022 in Korea.

Supplement 10. Trends in the sex ratio of alcohol-related mortality, 2012–2022 in Korea.

Supplement 11. Trends in mortality rates due to dementia by cause, 2012–2022 in Korea.

Supplement 12. COVID-19-related deaths per 100,000 population by age and sex, 2021 and 2022 in Korea.

References

1. Shin HY, Lee JY, Song J, Lee S, Lee J, Lim B, Kim H, Huh S. Cause-of-death statistics in the Republic of Korea, 2014. J Korean Med Assoc 2016;59:221-232. https://doi.org/10.5124/jkma.2016.59.3.221
2. Vital Statistics Division; Statistics Korea; Shin HY, Lee JY, Kim JE, Lee S, Youn H, Kim H, Lee J, Park MS, Huh S. Cause-of-death statistics in 2016 in the Republic of Korea. J Korean Med Assoc 2018;61:573-584. https://doi.org/10.5124/jkma.2018.61.9.573
3. Vital Statistics Division; Statistics Korea; Shin HY, Kim J, Lee S, Park MS, Park S, Huh S. Cause-of-death statistics in 2018 in the Republic of Korea. J Korean Med Assoc 2020;63:286-297. https://doi.org/10.5124/jkma.2020.63.5.286
4. Vital Statistics Division; Statistics Korea; Noh H, Seo J, Lee S, Yi N, Park S, Huh S. Statistical analysis of the cause of death in Korea in 2019. J Korean Med Assoc 2022;65:748-757. https://doi.org/10.5124/jkma.2022.65.11.748
5. Vital Statistics Division; Statistics Korea; Noh H, Seo J, Lee S, Yi N, Park S, Choi YJ, Huh S. Cause-of-death statistics in 2020 in the Republic of Korea. J Korean Med Assoc 2023;66:132-142. https://doi.org/10.5124/jkma.2023.66.2.132
6. Oh JH, So J, Lee S, Lim Y. Cause-of-death statistics in Korea from 2021. J Korean Med Assoc Forthcoming 2023.
7. Statistics Korea. Population statistics [Internet]. Statistics Korea; c2025 [cited 2025 Jan 29]. Available from: http://kostat.go.kr/
8. Park AK, Kim IH, Lee CY, Kim JA, Lee H, Kim HM, Lee NJ, Woo S, Lee J, Rhee J, Yoo CK, Kim EJ. Rapid emergence of the Omicron variant of severe acute respiratory syndrome coronavirus 2 in Korea. Ann Lab Med 2023;43:211-213. https://doi.org/10.3343/alm.2023.43.2.211
9. Im SJ, Shin JY, Lee DH. Excess deaths in Korea during the COVID-19 pandemic: 2020-2022. J Prev Med Public Health 2024;57:480-489. https://doi.org/10.3961/jpmph.24.254
10. Park S, Suh YK. A comprehensive analysis of COVID-19 vaccine discourse by vaccine brand on Twitter in Korea: topic and sentiment analysis. J Med Internet Res 2023;25:e42623. https://doi.org/10.2196/42623
11. Shi HJ, Yang J, Eom JS, Ko JH, Peck KR, Kim UJ, Jung SI, Kim S, Seok H, Hyun M, Kim HA, Kim B, Joo EJ, Cheong HS, Jun CH, Wi YM, Kim J, Kym S, Lim S, Park Y. Clinical characteristics and risk factors for mortality in critical COVID-19 patients aged 50 years or younger during Omicron wave in Korea: comparison with patients older than 50 years of age. J Korean Med Sci 2023;38:e217. https://doi.org/10.3346/jkms.2023.38.e217
12. Cho YS, Lee SH, So HJ, Kim DW, Choi YJ, Jeon HH. Effect of gastric cancer screening on patients with gastric cancer: a nationwide population-based study. J Dig Cancer Res 2020;

8:102-108.

13. Lee HJ, Lee K, Kim BC, Jun JK, Choi KS, Suh M. Effectiveness of the Korean National Cancer Screening Program in reducing colorectal cancer mortality. Cancers (Basel) 2024;16:4278. https://doi.org/10.3390/cancers16244278

14. Kim EY, Kim YJ, Choi WJ, Jeon JS, Kim MY, Oh DH, Jin KN, Cho YJ. Concordance rate of radiologists and a commercialized deep-learning solution for chest X-ray: real-world experience with a multicenter health screening cohort. PLoS One 2022; 17:e0264383. https://doi.org/10.1371/journal.pone.0264383

eISSN: 2234-2591
Open Access

Guidelines

Ewha Med J 2025;48(3):e47
https://doi.org/10.12771/emj.2025.00696

The Ewha Medical Journal

Ten guidelines for a healthy life: Korean Medical Association Statement (2017)

Chul Min Ahn[1], Jeong-Ho Chae[2], Jung-Seok Choi[3], Yong Pil Chong[4], Byung Chul Chun[5], Eun Mi Chun[6], Bo Seung Kang[7], Dai Jin Kim[2], Yeol Kim[8], Jun Soo Kwon[9], Sang Haak Lee[10], Won-Chul Lee[11], Yu Jin Lee[12], Jong Han Leem[13], Soo Lim[14], Saejong Park[15], Dongwook Shin[16], Hyeon Woo Yim[11], Kwang Ha Yoo[17], Dae Hyun Yoon[18], Ho Joo Yoon[19]

[1]*Pulmonary Division, Department of Internal Medicine, Gangnam Severance Hospital, Yonsei University College of Medicine, Seoul, Korea*
[2]*Department of Psychiatry, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul, Korea*
[3]*Department of Psychiatry, SMG-SNU Boramae Medical Center, Seoul National University College of Medicine, Seoul, Korea*
[4]*Department of Infectious Diseases, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea*
[5]*Department of Preventive Medicine, Korea University College of Medicine, Seoul, Korea*
[6]*Division of Pulmonary, Critical Care Medicine, Department of Internal Medicine, Ewha Womans University College of Medicine, Seoul, Korea*
[7]*Department of Emergency Medicine, Hanyang University Guri Hospital, Guri, Korea*
[8]*Graduate School of Cancer Science and Policy, National Cancer Center, Goyang, Korea*
[9]*Department of Psychiatry, Seoul National University Hospital, Seoul, Korea*
[10]*Division of Pulmonary, Critical Care, and Sleep Medicine, Department of Internal Medicine, College of Medicine, The Catholic University of Korea, Seoul, Korea*
[11]*Department of Preventive Medicine, College of Medicine, The Catholic University of Korea, Seoul, Korea*
[12]*Department of Neuropsychiatry, Seoul National University Hospital, Seoul, Korea*
[13]*Department of Occupational and Environmental Medicine, Inha University Hospital, Incheon, Korea*
[14]*Department of Internal Medicine, Seoul National University Bundang Hospital, Seoul National University College of Medicine, Seongnam, Korea*
[15]*Korea Institute of Sports Science, Seoul, Korea*
[16]*Department of Family Medicine, Samsung Medical Center, Seoul, Korea*
[17]*Division of Pulmonology and Allergy, Department of Internal Medicine, Konkuk University School of Medicine, Seoul, Korea*
[18]*Department of Neuropsychiatry, Healthcare System Gangnam Center, Seoul National University Hospital, Seoul, Korea*
[19]*Division of Allergy and Pulmonary Medicine, Department of Internal Medicine, Hanyang University Seoul Hospital, Seoul, Korea*

Quit smoking

Smoking remains a leading cause of premature death due to cardiovascular disease, chronic obstructive pulmonary disease, and lung cancer. Its harmful effects extend beyond smokers to nonsmokers, as secondhand smoke significantly endangers others—especially children, who face increased risks of respiratory diseases and asthma [1]. Although more than 70% of smokers express a desire to quit, nicotine dependence is recognized as a chronic, relapsing disease. As a result, unassisted quit attempts typically yield a very low success rate of only 3%–5% [2]. Therefore, effective, evidence-based interventions are essential for successful smoking cessation.

A comprehensive, multi-faceted approach greatly improves cessation outcomes. One of the most impactful initial steps is to publicly announce your decision to quit to family, friends, and colleagues. Smoking behaviors are strongly influenced by social networks; having a spouse or friend who smokes can reduce the likelihood of quitting, while support from nonsmokers can significantly increase quit rates.

*Corresponding email: Contact the Korean Medical Association at http://kma.org/english/contact.asp

Received: July 7, 2025 Accepted: July 22, 2025

This is an abridged article based on the "Ten guidelines for a healthy life: Korean Medical Association statement (2017)," published by the Korean Medical Association in 2017. Original Korean version is available at Supplement 1 and the English version is available at Supplement 2. This article is shared with the permission of the Korean Medical Association to make its content accessible not only to Koreans but also to people worldwide for the promotion of health.

Accessing professional medical support is also crucial. Even brief consultations with specialists can increase quit rates, and the benefit increases with more frequent and longer counseling. Clinicians commonly utilize the "5 A's" framework (ask, advise, assess, assist, arrange) to guide smokers through the cessation process. For individuals not yet ready to quit, the "5 R's" framework (relevance, risks, rewards, roadblocks, repetition) is used to build motivation. The combination of counseling and pharmacotherapy has consistently proven to be the most effective pathway to quitting [2].

Pharmacotherapy is strongly recommended for nearly all smokers seeking to quit, as it plays a vital role in managing nicotine addiction. First-line, Food and Drug Administration-approved medications include nicotine replacement therapy, bupropion, and varenicline [1-3] (Table 1). These medications each target withdrawal symptoms and cravings in different ways, and extensive studies have established their safety and efficacy, including among patients with stable psychiatric disorders [4]. In some trials, varenicline has demonstrated greater efficacy than bupropion, though it may also affect an individual's response to alcohol.

Effectively managing withdrawal symptoms is key to long-term success. Weight gain after quitting is common but can be controlled through regular exercise, a healthy diet, and stress management strategies. These not only help maintain a healthy weight but also reduce relapse triggers such as anxiety, supporting sustained abstinence. In contrast, electronic cigarettes are not recommended as cessation aids, as they contain toxic chemicals and lack sufficient safety data [5]. By integrating these evidence-based strategies, smokers can maximize their chances for lasting cessation.

Drink in moderation

The "alcohol flush reaction," characterized by facial redness, rapid heart rate, and nausea following alcohol consumption, is an important but frequently overlooked health indicator in social situations. Contrary to common belief, this reaction is not simply a sign of low tolerance that can be overcome with practice; rather, it signifies a genetic inability to safely metabolize alcohol. Ethanol, the intoxicating component of alcoholic beverages, is first converted into acetaldehyde, a toxic substance. The enzyme aldehyde dehydrogenase (ALDH) then breaks acetaldehyde down into harmless acetate. However, many individuals inherit a less effective form of the ALDH enzyme, due to genetic factors [6].

This condition, known as ALDH deficiency, results in the rapid and hazardous buildup of acetaldehyde even with small amounts of alcohol. It is particularly prevalent among East Asians, affecting approximately 30% of Koreans, 30%–33% of Chinese, and 45% of Japanese people. In contrast, ALDH deficiency is rare among European, North American, and African populations [7]. Acetaldehyde is classified as a Group 1 carcinogen by the World Health Organization (WHO). Therefore, pressuring someone with alcohol flush reaction to drink is equivalent to encouraging the consumption of a known cancer-causing substance, which significantly raises the risk of esophageal, head, and neck cancers. Social practices that promote drinking, especially in professional settings, pose serious risks for these individuals. A responsible approach is to fully exempt them from drinking.

Even for those without this genetic sensitivity, moderation remains essential to prevent health risks. An individual's ability to metabolize alcohol depends on various factors, including gender, body size, and overall health, but following general guidelines can help minimize harm. For occasional drinkers (1–3 times per month), a safe limit is usually 3–4 standard drinks for men and 2–3 for women, consumed slowly over 2–3 hours and accompanied by food and water [8].

Regular, frequent drinking, even at moderate amounts per occasion, increases the risk of mouth, throat, liver, and breast cancers. To reduce these risks, men who drink several times a week should limit their total weekly intake to the equivalent of 2 bottles of soju (or 8 cans of beer), while women should not exceed half that amount [9]. One practical and effective way to reduce both

Table 1. First-line agents for tobacco cessation that have been approved by the Food and Drug Administration

Treatment	Mechanism of action	Common side effects & considerations
Nicotine replacement therapy	Delivers nicotine without harmful toxins to reduce physiological withdrawal symptoms. Available as patches, gum, and lozenges [2].	Skin irritation (patches), nausea, tachycardia. Can be used by those with stable cardiovascular disease.
Sustained-release bupropion	Norepinephrine-dopamine reuptake inhibitor (antidepressant) that reduces cravings and withdrawal. May help control weight gain [1].	Insomnia, dry mouth, headache. A daily dose of 300 mg should not be exceeded due to a rare risk of seizure.
Varenicline	A partial agonist that binds to nicotine receptors to reduce withdrawal symptoms while also blocking the rewarding effects of nicotine [3].	Nausea, insomnia, nightmares. Patients should reduce alcohol intake and be aware of a rare seizure risk.

overall alcohol consumption and cancer risk is to schedule several alcohol-free days each week [10]. Understanding these health risks is essential for fostering a safer and healthier drinking culture.

Eat a balanced diet

Maintaining a balanced diet is fundamental to the prevention of chronic diseases, particularly as dietary habits in Korea increasingly shift toward westernized and unbalanced patterns. Achieving a healthy lifestyle requires a deliberate effort to regulate nutrient intake, control sugar consumption, and maintain a healthy weight.

Balance your macronutrients

For optimal health, the Korean Dietary Reference Intakes recommend a macronutrient distribution of 55%–65% carbohydrates, 7%–20% protein, and 15%–30% fat [11]. This guidance is especially important as many older Koreans tend to consume excessive carbohydrates while lacking sufficient fat in their diets [11]. Striking the right balance through whole foods is crucial, while unscientific fad diets should be avoided. For example, the low-carb high-fat diet is considered an extreme and unsustainable approach. By drastically reducing carbohydrates and promoting fat intake above 70%, it can dangerously elevate low-density lipoprotein ("bad") cholesterol, cause micronutrient deficiencies, and lead to adverse effects such as poor concentration due to ketosis [12]. Individuals with chronic diseases like diabetes or cardiovascular conditions must consult a physician before making significant dietary changes, as abrupt shifts may interfere with medications and worsen their health [12].

Reduce added sugar

The WHO recommends limiting added sugar to less than 10% of total daily calories—equivalent to about 50 grams for a 2,000-kcal diet [13]. Sugar intake in Korea continues to rise, mainly due to increased consumption of processed foods and beverages [11]. The main sources of added sugar differ by age group: children and adolescents primarily consume it through sodas, while adults over 30 often get it from sweetened coffee [11]. The most effective way to reduce sugar intake is to replace sugar-sweetened drinks with water and to remain vigilant about hidden sugars in juices, snacks, and processed milk.

Maintain a healthy weight

Rising obesity rates in Korea are closely linked to high-calorie diets and declining physical activity. Obesity—especially increased abdominal fat—is a significant contributor to metabolic syndrome, which raises the risk for type 2 diabetes, hypertension, and cardiovascular disease. This issue is particularly alarming among adolescents, as the prevalence of metabolic syndrome in Korean youth has doubled over a decade, while rates have declined in the United States [14]. This highlights the critical importance of establishing healthy habits in childhood. Furthermore, rapid weight gain in early adulthood is a strong predictor of future coronary artery disease [15]. To prevent chronic illness, it is essential to maintain a healthy weight throughout life by balancing caloric intake with regular physical activity.

Be physically active

In today's industrialized society, sedentary lifestyles have become a major health hazard. The WHO recognizes physical inactivity as 1 of the 4 leading risk factors for global mortality, along with high blood pressure, smoking, and hyperglycemia [16]. Physical inactivity is a principal driver of chronic diseases and is estimated to cause 21%–25% of breast and colon cancers, 27% of diabetes cases, and 30% of ischemic heart disease [16]. Engaging in regular physical activity is essential not only for preventing these conditions but also for enhancing both physical and mental well-being.

To counteract sedentary habits, physical activity should be seamlessly incorporated into daily life. For busy individuals, this means seizing opportunities for movement during routine tasks, such as taking the stairs instead of the elevator, walking or cycling for short trips, or performing household chores with greater energy [17]. Breaking up long periods of sitting is equally important. Studies have shown that watching TV for more than 2 hours per day increases the risk of diabetes and cardiovascular disease, and that prolonged sedentary time elevates mortality risk regardless of exercise habits. The basic rule is simple and clear: move more, sit less.

For structured exercise, current guidelines recommend a combination of aerobic and strength-training activities. Adults should aim for at least 150 minutes of moderate-intensity aerobic activity (such as brisk walking) or 75 minutes of vigorous-intensity activity each week, performed in sessions of at least 10 minutes. This improves cardiorespiratory fitness, which is inversely related to metabolic disease and mortality [18].

In addition, muscle-strengthening activities involving all major muscle groups should be performed on 2 or more days per week [19]. Resistance training helps build muscle mass, raises basal metabolic rate, and improves blood sugar control, thereby supporting weight management. These principles can be adapted for all ages; older adults should also include balance exercises to prevent falls [19].

The benefits of physical activity extend well beyond disease prevention. An active lifestyle improves bone health, alleviates symptoms of depression and anxiety, boosts mood and self-esteem, and contributes to a higher overall quality of life [20]. Physical activity remains one of the most accessible and effective tools for promoting public health.

Have a regular sleep schedule

Sound sleep is an essential pillar of both physical and mental health. Chronic sleep deprivation impairs judgment and mood, increases the risk of accidents, and, over time, raises the likelihood of developing obesity, diabetes, and cardiovascular disease. Adhering to principles of good sleep hygiene can markedly improve sleep quality and enhance overall health and vitality.

A consistent daily routine is the cornerstone of healthy sleep, reinforcing the body's natural 24-hour biological clock [21]. The most important rule is to wake up at the same time every day, including weekends, regardless of when you go to bed. To strengthen the link between bed and sleep, only go to bed when you genuinely feel sleepy. If you remain awake after a reasonable period, get up and engage in a calming activity until drowsiness returns [21]. While short daytime naps can be rejuvenating, they reduce the "homeostatic sleep drive" needed for nighttime rest; if you must nap, limit it to under 30 minutes to avoid disrupting your main sleep period [22].

Daytime lifestyle choices significantly influence sleep at night. Regular daytime exercise is especially beneficial, as it has been shown to increase sleep duration, shorten the time needed to fall asleep, and improve overall sleep quality. However, vigorous exercise should be avoided in the few hours before bedtime, as it can be overstimulating and delay sleep onset [23].

Caffeine and alcohol must also be managed carefully. Caffeine can substantially reduce total sleep time and disrupt sleep cycles, with adverse effects persisting even when consumed up to 6 hours before bedtime [23]. Nicotine is another stimulant that impairs sleep quality. Although many use alcohol to relax, it initially induces drowsiness but ultimately fragments sleep, diminishes its restorative effects, and causes frequent awakenings during the night [24].

By maintaining a regular sleep schedule, ensuring sufficient rest for your age, and making mindful choices about exercise, caffeine, and alcohol, you can proactively support your sleep and foster a healthier, more energetic life.

Think positively

A happy and meaningful life is not left to chance but is a skill that can be cultivated through conscious practice. According to the principles of positive psychology, intentionally fostering positive emotions can expand our thinking, improve physical health, and strengthen our connections with others [25]. By developing habits focused on gratitude, self-worth, and relationships, we can significantly enhance our well-being.

The first step is to appreciate the small things in life. Research consistently shows that people who regularly practice gratitude experience more joy, optimism, and energy [26]. This involves deliberately noticing the good in everyday experiences, no matter how minor. At the end of each day, take a moment to reflect on what went well. Savoring positive moments and expressing thanks—whether by writing them down or telling someone directly—is a powerful intervention that can lead to lasting increases in happiness [25]. This simple habit shifts your focus away from what is lacking and toward what is abundant.

Second, resist the urge to compare yourself to others. Instead, focus on your personal growth and unique character strengths. Positive psychology teaches that human goodness and excellence are just as real and important as disease and disorder [25]. True confidence comes not from feeling superior to someone else, but from recognizing your own progress compared to yesterday. Identifying and using your personal strengths is a skill that can be learned, fostering resilience and helping you navigate challenges more happily.

Finally, remember that happiness is rooted in strong relationships. The need to belong and form positive attachments is a fundamental human motivation [27]. People with robust social support networks are not only happier but also physically healthier and better equipped to manage life's inevitable stressors [28]. Happiness also grows from engaging in valued activities with those you care about [29]. Prioritize meals and pleasant conversations with friends and family. Investing in these connections is one of the most dependable and rewarding ways to achieve a fulfilling life.

Receive routine health screenings and immunizations

Prevention is always better than cure. South Korea's national health screening and immunization programs are essential for proactive health management, helping to detect diseases early and prevent infectious outbreaks. Making use of these programs is a cornerstone of maintaining a long and healthy life.

Health screenings: your first line of defense

Regular health screenings are crucial because many chronic diseases and cancers are asymptomatic in their early, most treatable stages. Early detection through screening significantly improves outcomes. For example, screenings can reduce mortality rates by 25% for breast cancer, 20% for colon cancer, and 42% for cardiovascular diseases [30].

South Korea offers a comprehensive, life-cycle-based National Health Screening Program, which includes a targeted National Cancer Screening Program for 5 major cancers: stomach, liver, colon, breast, and cervix. Despite the proven benefits and low cost of these programs, participation rates remain suboptimal, with only 48.3% of eligible individuals receiving cancer screenings [31]. An even more critical issue is the low follow-up rate; many people with a positive or suspicious result do not complete the necessary confirmatory exams. It is important to remember that screening is intended to identify potential problems. Therefore, always consult a doctor for follow-up if you receive an abnormal result. Following the recommended screening schedule and acting on the results are the best ways to safeguard your health.

Immunizations: protecting yourself and your community

Immunizations are the single most effective way to prevent infectious diseases. They protect not only the individual but also the community by establishing herd immunity [32]. Korea's national hepatitis B vaccination program, which dramatically reduced the rates of chronic hepatitis and liver cancer, is a powerful example of the public health impact of widespread immunization.

Korea's National Immunization Program provides essential vaccines for children and adults (Tables 2, 3) [11,12,33,34]. While childhood vaccination coverage is high, immunization rates among at-risk adults, such as for influenza, remain concerningly low. Vaccines are extremely safe and effective, and the tremendous benefits of disease prevention far outweigh the minimal risk of ad-

Table 2. Recommended immunization schedule for children

Vaccine	Age/doses
BCG (tuberculosis)	Within 4 weeks of birth (1 dose)
HepB (hepatitis B)	At birth, 1 month, 6 months (3 doses)
DTaP (diphtheria, tetanus, pertussis)	2, 4, 6, 15–18 months; 4–6 years (5 doses)
Tdap/Td	11–12 years (1 booster dose)
IPV (polio)	2, 4, 6–18 months; 4–6 years (4 doses)
Hib (Haemophilus influenzae type b)	2, 4, 6, 12–15 months (4 doses)
PCV (pneumococcal conjugate)	2, 4, 6, 12–15 months (4 doses)
MMR (measles, mumps, rubella)	12–15 months, 4–6 years (2 doses)
Varicella (chickenpox)	12–15 months (1 dose)
HepA (hepatitis A)	12–23 months (2 doses, 6 months apart)
Japanese Encephalitis	Varies by vaccine type (2 or 5 doses)
HPV (human papillomavirus)	11–12 years (2 doses, 6 months apart)
Influenza (flu)	Annually, starting at 6 months

Adapted from Korea Centers for Disease Control and Prevention [11]. This is a simplified schedule; consult a physician for personalized advice.

Table 3. Recommended immunization schedule for adults (19+)

Vaccine	Recommendation
Influenza (flu)	Annually for all adults, especially those ≥ 65 or with chronic conditions.
Tdap/Td (tetanus, diphtheria, pertussis)	Td booster every 10 years. Substitute 1 Td booster with Tdap once.
Pneumococcal (PPSV23)	1 dose for all adults ≥ 65. 1–2 doses for adults 19–64 with certain chronic health conditions (e.g., heart/lung disease, diabetes, smokers).
Herpes zoster (shingles)	1 dose recommended for all adults ≥ 60.
Hepatitis A	2 doses for at-risk adults (e.g., chronic liver disease) or those seeking protection.
Hepatitis B	3 doses for at-risk adults (e.g., chronic liver disease, diabetes) who were not vaccinated as children.
MMR (measles, mumps, rubella)	1–2 doses for adults born in 1957 or later without evidence of immunity.
Varicella (chickenpox)	2 doses for adults without evidence of immunity.

Adapted from Korea Centers for Disease Control and Prevention [12]. Recommendations may vary based on health status, occupation, and travel. Consult a physician.

verse reactions. Following the recommended immunization schedule is essential for your health and the well-being of the community.

Manage stress

While some stress can provide energy and motivation, excessive stress is a primary cause of physical and mental illness. Since modern life makes stress unavoidable, the key is not to eliminate it but to manage it effectively. Mastering stress requires a combination of shifting your mindset, developing active coping strategies, and engaging in restorative leisure activities.

The power of your mindset

The intensity of stress you experience is determined not by the situation itself, but by your thoughts about it. By recognizing and challenging overly negative or irrational thoughts, you can fundamentally transform your emotional responses and reduce the harmful effects of stress [35].

This mental resilience is anchored by strong self-esteem. Self-esteem, which is rooted in self-acceptance and a positive view of your abilities, acts as a crucial buffer against life's challenges. It empowers you to handle criticism and setbacks without losing courage or initiative. Rather than trying to avoid stress, building robust self-esteem enables you to accept stress as a natural part of life. You can nurture this by acknowledging your imperfections while still valuing your unique strengths, and by regularly stepping away from daily pressures to reflect and regain a positive perspective.

Active coping strategies

When stress arises, you can calm your body's physical and mental responses with practical techniques. Meditation and mindful breathing are proven methods for soothing the nervous system. Even just 10 minutes a day can lower blood pressure, reduce anxiety, and improve focus. When managing anger, it is best to pause and observe your feelings before reacting. If you must express anger, focus on communicating the specific behavior that upset you rather than making general accusations.

Excessive anxiety can also be managed by engaging in activities that signal safety to the brain. These include taking short walks in nature and having warm, empathetic conversations with trusted friends, which help shift your mind from a state of high alert to one of peace and connection [36].

The vitality of leisure

Leisure is not an indulgence; it is an essential lubricant that keeps life running smoothly. Regularly engaging in a hobby—whether physical, social, cultural, or creative—regenerates energy, prevents burnout, and provides emotional stability [37-39]. The specific activity is less important than the enjoyment and immersion it brings. Hobbies offer a healthy, positive alternative to negative coping mechanisms such as drinking or smoking. By making time for activities that bring you joy and allow you to disconnect from daily pressures, you actively recharge your mind and build a more resilient, vibrant life.

Pay attention to particulate matter and emerging infectious diseases

Modern life presents 2 growing threats to public health: chronic exposure to particulate matter (PM) in air pollution and the acute risk of emerging infectious diseases. Addressing these challenges requires both government action and proactive personal habits to safeguard individual and community well-being.

The invisible danger of particulate matter

Particulate matter—especially fine particles ($PM_{2.5}$) from traffic and industrial sources—poses a severe health risk. These tiny particles can bypass the body's natural defenses, penetrate deep into the lungs, and enter the bloodstream [40]. They often carry toxic heavy metals and other harmful substances. The WHO has classified outdoor air pollution and diesel engine exhaust as Group 1 carcinogens, definitively linking them to cancer [41].

Long-term exposure significantly increases mortality from cardiovascular diseases such as heart attacks and strokes, as well as from respiratory illnesses and lung cancer [42]. The Organization for Economic Cooperation and Development has issued a stark warning: without significant change, South Korea faces the highest projected mortality rate and economic loss from air pollution among all member nations by 2060 [43]. To reduce this risk, it is crucial to limit exposure by staying indoors and wearing a certified mask when PM levels are high and to reduce emissions by minimizing personal vehicle use.

Preventing emerging infectious diseases

In today's interconnected world, infectious diseases can rapidly spread across borders. Preventing their transmission relies on 2 key strategies: travel preparation and daily health etiquette.

Before traveling abroad, research the health risks at your destination. Visit an infectious disease or travel clinic at least 1 month before your trip. This ensures time for any necessary vaccinations [13] (Table 4), as immunity may take several weeks to develop, and for obtaining preventive medications such as those for malaria.

Table 4. Recommended immunizations for overseas travelers

Vaccine	Recommended for	Vaccination notes
Yellow fever	Travelers to endemic regions in Africa & South America. Required for entry to some countries.	1 dose every 10 years. Must be given at least 10 days before arrival.
Hepatitis A	All travelers to developing countries, especially those under 30 without immunity.	2 doses, given 6–12 months apart.
Typhoid	Travelers to South Asia, Southeast Asia, and other areas with poor sanitation.	1 dose, provides protection for approximately 2 years.
Meningococcal	Travelers to the “meningitis belt” in sub-Saharan Africa or for the Hajj pilgrimage.	1 dose. Revaccination may be needed after 5 years.
Rabies	Long-term travelers, veterinarians, or those working with animals in high-risk areas.	3 doses (pre-exposure prophylaxis).
Japanese encephalitis	Travelers spending extended time in rural or agricultural areas of Asia.	2 or 5 doses depending on vaccine type.
Cholera	Aid and refugee workers in areas with active outbreaks.	Oral vaccine (Dukoral).

Adapted from the Korean Society of Infectious Diseases [13]. This is a general guide; always consult a travel medicine specialist for personalized recommendations based on your specific itinerary and health status.

At home and while traveling, practicing basic “health etiquette” is your first line of defense. Proper hand hygiene is one of the most effective ways to prevent illness, stopping up to half of all respiratory and diarrheal infections [44]. Additionally, practice cough etiquette: cover your mouth and nose with your sleeve (not your hand), wear a mask when sick, and avoid crowded places to protect others. These simple habits are a foundation of public health.

Avoid excessive exposure to mobile devices

While smart devices are essential tools in modern life, excessive use poses significant risks to both physical and mental health. Mindful usage—knowing when, where, and how to disconnect—is crucial for preventing the negative consequences of digital overexposure, including obesity, sleep disruption, and impaired child development.

Disconnect during meals to support healthy habits

Using smartphones during meals is strongly linked to poor dietary choices and weight gain. Adolescents who spend over 5 hours a day on screens are twice as likely to consume sugary drinks and 43% more likely to be obese [45]. This is due to several factors: screen time is a sedentary activity, it increases exposure to junk food advertising, and distracted eating often results in overeating. In addition, blue-enriched light from LED (light-emitting diode) screens can disrupt insulin metabolism and blood sugar levels, further contributing to metabolic issues [46]. To encourage healthier eating, it is essential to put devices away and practice mindful eating without digital distractions.

Power down before bed for better sleep

The blue light emitted from smartphones, tablets, and computers is a major disruptor of the body’s natural sleep-wake cycle. Exposure within 2 hours of bedtime suppresses melatonin production—the hormone that signals the brain to sleep [47]. This can make it harder to fall asleep, lower sleep quality, and cause next-day fatigue, impairing concentration and performance at school or work. To promote better sleep, establish a strict, screen-free period of at least 2 hours before bedtime. Your bedroom should be a sanctuary for sleep, not a place for late-night scrolling.

Protect children’s development from excessive screen time

For infants and toddlers, exposure to smart devices is especially harmful. The brain undergoes its most rapid development from birth to age 2, a process that depends on active, real-world interaction—not passive screen time [48]. Early and excessive exposure can impede cognitive development, language acquisition, and social skills [49]. The risk of smartphone overdependence is rising alarmingly rapidly in this age group.

Parents must set a strong example, as their own screen habits greatly influence their children’s. It is critical to severely restrict—or ideally, eliminate—screen time for children under 2. For older children, parents should set clear boundaries and encourage face-to-face play and group activities to support healthy physical, cognitive, and social development.

ORCID

Chul Min Ahn: https://orcid.org/0000-0002-7071-4370
Jeong-Ho Chae: https://orcid.org/0000-0002-6070-9324
Jung-Seok Choi: https://orcid.org/0000-0003-2139-0522
Yong Pil Chong: https://orcid.org/0000-0003-1672-3185
Byung Chul Chun: https://orcid.org/0000-0001-6576-8916

Eun Mi Chun: https://orcid.org/0000-0001-9616-2722
Bo Seung Kang: https://orcid.org/0000-0002-0792-0198
Dai Jin Kim: https://orcid.org/0000-0001-9408-5639
Yeol Kim: https://orcid.org/0000-0003-1142-1559
Jun Soo Kwon: https://orcid.org/0000-0002-1060-1462
Sang Haak Lee: https://orcid.org/0000-0001-6259-7656
Won-Chul Lee: https://orcid.org/0000-0002-5483-1614
Yu Jin Lee: https://orcid.org/0000-0001-5195-2579
Jong Han Leem: https://orcid.org/0000-0003-3292-6492
Soo Lim: https://orcid.org/0000-0002-4137-1671
Saejong Park: https://orcid.org/0000-0001-7229-5790
Dongwook Shin: https://orcid.org/0000-0001-8128-8920
Hyeon Woo Yim: https://orcid.org/0000-0002-3646-8161
Kwang Ha Yoo: https://orcid.org/0000-0001-9969-2657
Dae Hyun Yoon: https://orcid.org/0000-0003-4724-5980
Ho Joo Yoon: https://orcid.org/0000-0002-4645-4863

Authors' contributions

All authors contributed their chapters and did all work.

Conflict of interest

No potential conflict of interest relevant to this article was reported.

Funding

This guideline project was supported by the Korean Medical Association from 2016 to 2017.

Data availability

Not applicable.

Acknowledgments

Following physicians and societies contributed to the revision and final confirmation of the content:

Physicians: Se Young Choung: Department of Preventive Pharmacy and Toxicology, College of Pharmacy, Kyung Hee University, Seoul, Korea; Young-Chul Jung: Department of Psychiatry, Severance Hospital, Yonsei University College of Medicine, Seoul, Korea; Jongsung Kim: Department of Family Medicine, Chungnam National University Hospital, Daejeon, Korea; Jong Chan Lee: a medical specialist in gastroenterology in the Department of Internal Medicine, Seoul National University Bundang Hospital, Seongnam, Korea; Jung Eun Lee: Department of Food and Nutrition, Seoul National University, Seoul, Korea; Kwang Shik Park: KBS News medical correspondent, Seoul, Korea; Kyung Taek Rim: Occupational Safety and Health Research Institute, Korea Occupational Safety and Health Agency, Ulsan, Korea; Cheol Min Shin: Department of Internal Medicine, Seoul National University Bundang Hospital, Seongnam, Korea; Mi-Kyung Sung: Department of Food and Nutrition, Sookmyung Women's University, Seongnam, Korea; Hye Ran Yang: Department of Pediatrics, Seoul National University Bundang Hospital, Seongnam, Korea; Yong Kyu You: Nammoon Medical Clinic

Societies: Korea Occupational Safety & Health Agency; Korean Academy of Addiction Psychiatry; The Korean Academy of Family Medicine; The Korean Academy of Sleep Medicine (KASMED); The Korean Academy of Tuberculosis and Respiratory Diseases; Korean Association of Psychoanalysis (KAPA): Study Group and Allied Center of the IPA; Korean Diabetes Association (KDA); Korean Endocrine Society (KES); Korean Medical Practitioners Association; Korean Neuropsychiatric Association (KNPA); The Korean Nutrition Society; The Korean Society for Preventive Medicine (KSPM); The Korean Society for Research on Nicotine and Tobacco; Korean Society for Study of Obesity (KSSO); Korean Society of ALDH; Korean Society of Cancer Prevention (KSCP); The Korean Society of Gastroenterology; The Korean Society of Infectious Diseases; The Korean Society of Medical Informatics (KOSMI); The Korean Society of Neurogastroenterology and Motility; Korean Society of Occupational & Environmental Medicine; The Korean Society of Pediatric Gastroenterology, Hepatology and Nutrition (KSPGHAN); The Korean Vaccine Society

Supplementary materials

Supplementary files are available from Harvard Dataverse: https://doi.org/10.7910/DVN/JQ51SA

Supplement 1. Korean version of the "Ten guidelines for a healthy life: Korean Medical Association Statement (2017)."

Supplement 2. English version of "Ten guidelines for a healthy life: Korean Medical Association Statement (2017)."

References

1. European Network for Smoking and Tobacco Prevention aisbl (ENSP). ENSP Guidelines for treating tobacco dependence [Internet]. ENSP; 2016 [cited 2017 May 4]. Available from: https://ensp.network/wp-content/uploads/2021/01/English_Guidelines_2016.pdf
2. Shiffman S, Brockwell SE, Pillitteri JL, Gitchell JG. Use of smoking-cessation treatments in the United States. Am J Prev Med 2008;34:102-111. https://doi.org/10.1016/j.amepre.2007.09.033
3. Jorenby DE, Hays JT, Rigotti NA, Azoulay S, Watsky EJ, Williams KE, Billing CB, Gong J, Reeves KR. Efficacy of vareni-

cline, an alpha4beta2 nicotinic acetylcholine receptor partial agonist, vs placebo or sustained-release bupropion for smoking cessation: a randomized controlled trial. JAMA 2006;296:56-63. https://doi.org/10.1001/jama.296.1.56

4. Anthenelli RM, Benowitz NL, West R, St Aubin L, McRae T, Lawrence D, Ascher J, Russ C, Krishen A, Evins AE. Neuropsychiatric safety and efficacy of varenicline, bupropion, and nicotine patch in smokers with and without psychiatric disorders (EAGLES): a double-blind, randomised, placebo-controlled clinical trial. Lancet 2016;387:2507-2520. https://doi.org/10.1016/S0140-6736(16)30272-0
5. Adkison SE, O'Connor RJ, Bansal-Travers M, Hyland A, Borland R, Yong HH, Cummings KM, McNeill A, Thrasher JF, Hammond D, Fong GT. Electronic nicotine delivery systems: international tobacco control four-country survey. Am J Prev Med 2013;44:207-215. https://doi.org/10.1016/j.amepre.2012.10.018
6. Brooks PJ, Enoch MA, Goldman D, Li TK, Yokoyama A. The alcohol flushing response: an unrecognized risk factor for esophageal cancer from alcohol consumption. PLoS Med 2009;6:e50. https://doi.org/10.1371/journal.pmed.1000050
7. Eng MY, Luczak SE, Wall TL. ALDH2, ADH1B, and ADH1C genotypes in Asians: a literature review. Alcohol Res Health 2007;30:22-27.
8. Jung SY, Hong SP. Establishment of functionality evaluation system for hangover settlement of health functional food [Internet]. Ministry of Food and Drug Safety; 2003 [cited 2017 May 4]. Available from: https://scienceon.kisti.re.kr/commons/util/originalView.do?cn=TRKO200400000549&dbt=TRKO&rn=
9. Kim JS. Moderate alcohol consumption guidelines for Koreans. In: Proceedings of the Spring Conference of the Korean Academy of Family Medicine; 2015 Apr 17-18; Daejeon, Korea. The Korean Academy of Family Medicine;2015.
10. Department of Health. UK Chief Medical Officers' low risk drinking guidelines [Internet]. Department of Health; 2016 [cited 2017 May 1]. Available from: https://assets.publishing.service.gov.uk/media/5a80b7ed40f0b623026951db/UK_CMOs__report.pdf
11. Ministry of Health and Welfare. Government Establishes National Dietary Guidelines for Koreans [Internet]. Ministry of Health and Welfare; 2016 [cited 2016 Apr 4]. Available from: http://www.mohw.go.kr/react/al/sal0301vw.jsp?PAR_MENU_ID=04& MENU_ID=0403&CONT_SEQ=330959&page=1
12. Korean Endocrine Society; Korean Diabetes Association; Korean Society for the Study of Obesity; The Korean Nutrition Society; The Korean Society of Lipid and Atherosclerosis. Joint statement on the low-carb high-fat diet trend from 5 professional organizations [Internet]. The Korean Nutrition Society; 2016 [cited 2017 Apr 25]. Available from: http://www.kns.or.kr/News/Notice_View.asp?idx=670
13. World Health Organization. Guideline: sugars intake for adults and children [Internet]. World Health Organization; 2015 [cited 2017 Apr 25]. Available from: http://www.who.int/nutrition/publications/guidelines/sugars_intake/en/
14. Lim S, Jang HC, Park KS, Cho SI, Lee MG, Joung H, Mozumdar A, Liguori G. Changes in metabolic syndrome in American and Korean youth, 1997-2008. Pediatrics 2013;131:e214-e222. https://doi.org/10.1542/peds.2012-0761
15. Lim S, Choi SH, Kim KM, Choi SI, Chun EJ, Kim MJ, Park KS, Jang HC, Sattar N. The association of rate of weight gain during early adulthood with the prevalence of subclinical coronary artery disease in recently diagnosed type 2 diabetes: the MAXWEL-CAD study. Diabetes Care 2014;37:2491-2499. https://doi.org/10.2337/dc13-2365
16. World Health Organization. Global health risks: mortality and burden of disease attributable to selected major risks [Internet]. World Health Organization; 2009 [cited 2017 Apr 25]. Available from: http://www.who.int/iris/handle/10665/44203
17. World Health Organization. Global recommendations on physical activity for health [Internet]. World Health Organization; 2010 [cited 2017 Apr 25]. Available from: http://www.who.int/dietphysicalactivity/factsheet_recommendations/en/
18. Farrell SW, Kampert JB, Kohl HW, Barlow CE, Macera CA, Paffenbarger RS, Gibbons LW, Blair SN. Influences of cardiorespiratory fitness levels and other predictors on cardiovascular disease mortality in men. Med Sci Sports Exerc 1998;30:899-905. https://doi.org/10.1097/00005768-199806000-00019
19. Nelson ME, Rejeski WJ, Blair SN, Duncan PW, Judge JO, King AC, Macera CA, Castaneda-Sceppa C. Physical activity and public health in older adults: recommendation from the American College of Sports Medicine and the American Heart Association. Circulation 2007;116:1094-1105. https://doi.org/10.1161/CIRCULATIONAHA.107.185650
20. Lee CD, Blair SN, Jackson AS. Cardiorespiratory fitness, body composition, and all-cause and cardiovascular disease mortality in men. Am J Clin Nutr 1999;69:373-380. https://doi.org/10.1093/ajcn/69.3.373
21. Farrand P, Woodford J. Impact of support on the effectiveness of written cognitive behavioural self-help: a systematic review and meta-analysis of randomised controlled trials. Clin Psychol Rev 2013;33:182-195. https://doi.org/10.1016/j.cpr.2012.11.001

22. Czeisler CA, Allan JS, Strogatz SH, Ronda JM, Sanchez R, Rios CD, Freitag WO, Richardson GS, Kronauer RE. Bright light resets the human circadian pacemaker independent of the timing of the sleep-wake cycle. Science 1986;233:667-671. https://doi.org/10.1126/science.3726555

23. Irwin MR, Cole JC, Nicassio PM. Comparative meta-analysis of behavioral interventions for insomnia and their efficacy in middle-aged adults and in older adults 55+ years of age. Health Psychol 2006;25:3-14. https://doi.org/10.1037/0278-6133.25.1.3

24. Stein MD, Friedmann PD. Disturbed sleep and its relationship to alcohol use. Subst Abus 2005;26:1-13. https://doi.org/10.1300/j465v26n01_01

25. Seligman ME, Steen TA, Park N, Peterson C. Positive psychology progress: empirical validation of interventions. Am Psychol 2005;60:410-421. https://doi.org/10.1037/0003-066X.60.5.410

26. Emmons RA, McCullough ME. Counting blessings versus burdens: an experimental investigation of gratitude and subjective well-being in daily life. J Pers Soc Psychol 2003;84:377-389. https://doi.org/10.1037//0022-3514.84.2.377

27. Baumeister RF, Leary MR. The need to belong: desire for interpersonal attachments as a fundamental human motivation. Psychol Bull 1995;117:497-529. https://doi.org/10.1037/0033-2909.117.3.497

28. House JS, Landis KR, Umberson D. Social relationships and health. Science 1988;241:540-545. https://doi.org/10.1126/science.3399889

29. Myers DG, Diener E. Who is happy? Psychol Sci 1995;6:10-19. https://doi.org/10.1111/j.1467-9280.1995.tb00298.x

30. Lee H, Cho J, Shin DW, Lee SP, Hwang SS, Oh J, Yang HK, Hwang SH, Son KY, Chun SH, Cho B, Guallar E. Association of cardiovascular health screening with mortality, clinical outcomes, and health care cost: a nationwide cohort study. Prev Med 2015;70:19-25. https://doi.org/10.1016/j.ypmed.2014.11.007

31. National Health Insurance Service. 2015 National Health Screening Statistical Yearbook. National Health Insurance Service; 2016.

32. Centers for Disease Control and Prevention (CDC). Impact of vaccines universally recommended for children: United States, 1990-1998. MMWR Morb Mortal Wkly Rep 1999;48:243-248.

33. Korea Centers for Disease Control and Prevention; Korean Medical Association; Korea Advisory Committee on Immunization Practice. Standard immunization schedule (2017): Korea (for healthy children). Korea Centers for Disease Control and Prevention; 2017.

34. Korea Centers for Disease Control and Prevention. Epidemiology and management of vaccine preventable disease. 5th ed. Korea Centers for Disease Control and Prevention; 2017.

35. Fredrickson BL, Joiner T. Positive emotions trigger upward spirals toward emotional well-being. Psychol Sci 2002;13:172-175. https://doi.org/10.1111/1467-9280.00431

36. Howe D. Empathy: what it is and why it matters. JK Lee Forest of Knowledge; 2013.

37. Iso-Ahola SE. The social psychology of leisure and recreation. W. C. Brown Company Publisher; 1980.

38. Gwak HB. Leisure cultures. Daewangsa; 2005.

39. Kwon YC. An introduction to social education. Kyoyookbook; 1994.

40. Pope CA, Burnett RT, Thun MJ, Calle EE, Krewski D, Ito K, Thurston GD. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. JAMA 2002;287:1132-1141. https://doi.org/10.1001/jama.287.9.1132

41. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. Outdoor air pollution. IARC Monogr Eval Carcinog Risks Hum 2016;109:9-444.

42. World Health Organization. Burden of disease from ambient air pollution for 2012 [Internet]. World Health Organization; 2012 [cited 2017 Apr 23]. Available from: https://era.org.mt/wp-content/uploads/2019/05/Burden-of-disease-from-Ambient-Air-Pollution-for-2012.pdf

43. Organization for Economic Cooperation and Development (OECD). The economic consequences of outdoor air pollution [Internet]. OECD; 2016 [cited 2017 Apr 12]. Available from: https://www.oecd.org/content/dam/oecd/en/publications/reports/2016/06/the-economic-consequences-of-outdoor-air-pollution_g1g68583/9789264257474-en.pdf

44. Luby SP, Agboatwalla M, Feikin DR, Painter J, Billhimer W, Altaf A, Hoekstra RM. Effect of handwashing on child health: a randomised controlled trial. Lancet 2005;366:225-233. https://doi.org/10.1016/S0140-6736(05)66912-7

45. Kenney EL, Gortmaker SL. United States adolescents' television, computer, videogame, smartphone, and tablet use: associations with sugary drinks, sleep, physical activity, and obesity. J Pediatr 2017;182:144-149. https://doi.org/10.1016/j.jpeds.2016.11.015

46. Cheung IN, Zee PC, Shalman D, Malkani RG, Kang J, Reid KJ. Morning and evening blue-enriched light exposure alters metabolic function in normal weight adults. PLoS One 2016;11:e0155601. https://doi.org/10.1371/journal.pone.0155601

47. Wood B, Rea MS, Plitnick B, Figueiro MG. Light level and du-

ration of exposure determine the impact of self-luminous tablets on melatonin suppression. Appl Ergon 2013;44:237-240. https://doi.org/10.1016/j.apergo.2012.07.008

48. Dobbing J, Sands J. Quantitative growth and development of human brain. Arch Dis Child 1973;48:757-767. https://doi.org/10.1136/adc.48.10.757

49. Smetaniuk P. A preliminary investigation into the prevalence and prediction of problematic cell phone use. J Behav Addict 2014;3:41-53. https://doi.org/10.1556/JBA.3.2014.004

eISSN: 2234-2591
Open Access

Guidelines

Ewha Med J 2025;48(3):e48
https://doi.org/10.12771/emj.2025.00668

The Ewha Medical Journal

TRIPOD+AI 지침: 회귀 또는 머신러닝 방법을 사용하는 임상 예측모델 보고를 위한 최신 지침

TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods: a Korean translation

Gary S. Collins[1*], Karel G. M. Moons[2], Paula Dhiman[1], Richard D. Riley[3,4], Andrew L. Beam[5], Ben Van Calster[6,7], Marzyeh Ghassemi[8], Xiaoxuan Liu[9,10], Johannes B. Reitsma[2], Maarten van Smeden[2], Anne-Laure Boulesteix[11], Jennifer Catherine Camaradou[12,13], Leo Anthony Celi[14,15,16], Spiros Denaxas[17,18], Alastair K. Denniston[4,9], Ben Glocker[19], Robert M. Golub[20], Hugh Harvey[21], Georg Heinze[22], Michael M. Hoffman[23,24,25,26], André Pascal Kengne[27], Emily Lam[12], Naomi Lee[28], Elizabeth W. Loder[29,30], Lena Maier-Hein[31], Bilal A. Mateen[17,32,33], Melissa D. McCradden[34,35], Lauren Oakden-Rayner[36], Johan Ordish[37], Richard Parnell[12], Sherri Rose[38], Karandeep Singh[39], Laure Wynants[40], Patricia Logullo[1]

*For further information on the authors' affiliations, see **Additional information**.*

요약

최근 인공지능(artificial intelligence, AI) 방법, 특히 머신러닝의 발전에 따라 예측모델 개발에 대한 관심과 투자규모가 크게 증가하고 있다. 예측모델 연구가 실제 사용자에게 가치 있게 활용되기 위해서는, 연구자가 왜 연구를 수행했는지, 무엇을 했는지, 그리고 어떤 결과를 얻었는지를 투명하고 완전하며 정확하게 기술해야 한다. TRIPOD 지침의 개정판은 AI 방법을 적용한 예측모델 연구 전반을 일관성 있게 안내하고, 회귀분석이든 머신러닝이든 적용방법에 관계없이 모두를 아우르는 지침을 제공한다. TRIPOD+AI 지침은 27개 항목의 체크리스트, 각 항목별 보고 권고사항을 상세히 설명하는 확장 체크리스트, 그리고 13개 항목의 초록 전용 TRIPOD+AI 체크리스트로 구성된다. TRIPOD+AI의 목표는 저자가 연구를 완전하게 보고하도록 돕고, 동료 평가자와 편집자, 정책 입안자, 최종 사용자, 그리고 환자가 AI 기반 연구의 데이터, 방법, 결과 및 결론을 명확히 이해하도록 돕는 것이다. 이 권고안을 준수하면 연구에 소요되는 시간, 노력, 비용의 활용 효율성을 높일 수 있을 것이다.

서론

2015년에 발표된 TRIPOD (transparent reporting of a multi-

*Corresponding email: gary.collins@csm.ox.ac.uk
Received: July 17, 2025 Revised: July 30, 2025 Accepted: July 30, 2025

It is a Korean translation of the Collins GS. et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024;385:e078378. https://doi.org/10.1136/bmj-2023-078378. The translation was done with the permission of the TRIPOD Group. Korean medical terminology is based on the English-Korean Medical Terminology 6th edition, available at: https://term.kma.org/index.asp. Korean translation was done by Sun Huh (https://orcid.org/0000-0002-8559-8640), Hallym University. The Korean proofreading was conducted by YoonJoo Seo (https://orcid.org/0000-0002-0202-8352), InfoLumi, and the back-translation was performed by Jeong-Ju Yoo (https://orcid.org/0000-0002-7802-0381), Soonchunhyang University Bucheon Hospital. The back-translation was confirmed by Gary S. Collins (https://orcid.org/0000-0002-2772-2316), the first author of the original TRIPOD+AI statement.

variable prediction model for individual prognosis or diagnosis; 개인별 예후 또는 진단을 위한 다변량 예측모델의 투명한 보고) 지침은 예측모델의 개발 또는 성능 평가 연구에 대한 최소 보고 권고사항을 제시하였다. 이후 예측모델 분야에서는 머신러닝에 기반한 인공지능(artificial intelligence, AI) 기법의 보편화 등 다양한 방법론적 발전이 이루어져 예측모델 개발에 활용되고 있다. 이에 따라 TRIPOD 지침의 개정이 필요하게 되었다. TRIPOD+AI는 회귀분석이든 머신러닝 방법이든 상관없이 예측모델 연구의 보고를 위한 일관된 지침을 제공한다. 이 새로운 체크리스트는 TRIPOD 2015 체크리스트를 대체하는 것으로, 기존 버전은 더 이상 사용하지 말아야 한다. 이 논문에서는 TRIPOD+AI의 개발과정과 함께, 각 보고 권고사항에 대해 더 상세히 설명하는 27개 항목의 확장 체크리스트와 초록용 TRIPOD+AI 체크리스트를 제시한다. TRIPOD+AI의 목적은 예측모델의 개발 또는 성능 평가 연구에서 완전하고 정확하며 투명한 보고를 장려하는 것이다. 완전한 보고는 연구 평가, 모델의 평가 및 실제 적용을 촉진할 것이다.

예측모델은 다양한 의료환경에서 사용되며, 특정 결과값이나 위험도를 산출하는 데 활용된다. 대부분의 모델은 특정 건강상태(진단)의 존재 가능성이나 특정 결과가 미래에 발생할지(예후)를 예측한다[1]. 주요 용도는 임상적 의사결정 지원으로, 예를 들어 추가 검사가 필요한지, 질환 악화나 치료효과 모니터링, 치료 또는 생활습관 변화 시작 여부 결정 등에 활용된다. 널리 알려진 예측모델로는 EuroSCORE II(심장 수술)[2], Gail 모델(유방암)[3], Framingham 위험 점수(심혈관질환)[4], IMPACT(외상성 뇌손상)[5], FRAX(골다공증 및 고관절 골절)[6] 등이 있다.

예측모델은 생의학 문헌에서 매우 풍부하게 보고되고 있으며, 매년 수천 개의 모델이 출판되고 그 수는 점차 증가하고 있다[7,8]. 다양한 건강상태와 임상 결과에 대한 매우 많은 모델이 개발되었다. Coronavirus disease 2019 (COVID-19) 팬데믹 첫 1년 동안만 해도, 진단 및 예후 예측모델 연구가 최소 731편 발표되었다[9]. 이렇게 예측모델 개발에 대한 관심이 높음에도 불구하고, 이 분야에서는 보고의 투명성과 완전성, 그리고 그에 따른 활용 가능성에 대한 우려가 오래전부터 지속되어 왔다[10,11]. 보고가 불완전하거나 부정확하다면, 동료 평가자, 편집자, 의료인, 규제 당국, 환자, 그리고 일반 대중을 포함한 독자 입장에서는 연구설계와 방법을 비판적으로 평가하고 결과에 신뢰를 갖거나, 추가로 모델을 평가하거나 실제 적용하기가 어렵다. 모델에 대한 불충분한 보고로 설계나 데이터 수집, 연구 수행과정의 결함이 가려질 수 있으며, 이러한 모델이 임상경로에 실제로 적용되면 위해로 이어질 수 있다. 특히 편향을 줄이기 위한 충분한 조치가 마련되지 않은 경우 위해가 발생할 수 있다. 더 나은 보고는 신뢰를 제고하고, 예측모델의 의료현장 적용에 대한 환자 및 대중의 수용도에 긍정적인 영향을 미칠 수 있다. 연구자는 자신의 연구를 완전하고 투명하며 정직하게 보고할 윤리적·과학적 의무가 있다. Altman 등[12]이 언급한 바와 같이 "좋은 보고는 선택이 아니라 연구의 필수 요소"이며, 그렇지 못한 보고는 결국 불필요한 연구 낭비에 불과하다[13].

불완전한 보고에 대한 우려에 대응하기 위해[10,11,14,15], TRIPOD 지침이 2015년에 발표되어(TRIPOD 2015) 최소 보고 권고사항을 제시하였다[16,17]. TRIPOD 2015는 총 37개 항목의 체크리스트로 구성되어 있으며, 이 중 25개 항목은 개발 및 검증 연구 모두에 공통으로 적용되고, 모델 개발 연구에 6개, 검증 연구에 6개의 추가 항목이 있다. 또한 각 체크리스트 항목의 근거, 좋은 보고의 사례, 예측모델 연구의 설계·수행·분석 관련 논의 등을 담은 설명 및 해설 문서도 함께 제공된다[17]. TRIPOD 2015는 당대의 주류였던 회귀분석 기반 모델에 주로 초점을 맞추었다. 이후 예측모델 연구의 초록 보고(TRIPOD for Abstracts)[18], 군집화 데이터 사용 연구(TRIPOD-Cluster)[19,20], 예측모델 연구의 체계적 문헌고찰 및 메타분석(TRIPOD-SRMA)[21], 연구 프로토콜 준비 지침(TRIPOD-P)[22] 등 추가 지침이 개발되었다. 이러한 모든 지침과 별도 작성용 체크리스트 서식은 TRIPOD 웹사이트(https://www.tripod-statement.org/)에서 확인할 수 있다.

TRIPOD 2015 발표 이후, 예측모델링 분야에는 샘플 사이즈 산정 지침[23-27], 성능 평가방법[28-32], 공정성[33], 재현성[34], 오픈 사이언스 원칙 적용[35] 등 다양한 방법론적 발전이 이루어졌다. 이 중에서도 가장 큰 변화와 진보가 나타난 영역은 AI로 분류되는 기법의 발전에 기반한 분야이다. 데이터 접근성의 향상과 기성 머신러닝 소프트웨어의 보급으로 예측모델 개발은 훨씬 빠르고 쉬워졌다. 여러 임상환경과 광범위한 결과·건강상태에 대한 수많은 예측모델이 문헌에 보고되고 있으며, 동일 결과나 건강상태, 대상 집단에 대해 복수의 모델이 존재하는 경우도 많다[7,8,36]. 따라서 예측모델의 품질을 비판적으로 평가하고, 특정 환경이나 사용 목적에 적합한지 이해하는 능력이 더욱 중요해졌다. 이러한 능력은 완전하고 투명한 보고를 전제로 한다.

하지만 예측모델 연구를 평가한 체계적 문헌고찰에서는, 연구설계나 데이터 수집의 결함[37,38], 미흡한 방법론 사용[37,38], 핵심 세부 사항이 누락된 불완전한 보고[39-54], 그에 따른 높은 편향 위험[41,49,55-57], 오픈 사이언스 관행의 미준수[58], 과도한 해석이나 이른바 'spin'의 문제[59,60] 등이 자주 드러난다. 이러한 결함은 모델의 유용성과 안전성에 심각한 의문을 제기하며, 보건의료격차가 심화할 우려도 있다[61]. TRIPOD 2015는 모델링 방식이 중립적이며 보고 권고 대부분이 비회귀적 접근법에도 적용되지만, 머신러닝 기반 모델에는 추가적인 보고 고려사항이 필요하다. 예를 들어 회귀 기반 모델과 달리 머신러닝 모델은 구조의 유연성과 복잡성 때문에 단순 공식이 나오지 않거나, 사용된 예측변수 자체가 불명확한 경우가 많다. 이 때문에 기존 TRIPOD 2015에서는 다루지 않은 추가 보고사항이 필요하다. 방법론적 진전뿐 아니라, 공정성[62], 오픈 사이언스 관행의 확산[63], 환자 및 대중의 연구·실제 적용 참여 확대[64,65] 등도 함께 반영할 필요가 있다.

이 논문의 목적은 개정된 TRIPOD 지침의 개발과정을 설명하고, 새로운 TRIPOD+AI 체크리스트를 제시하며, 그 활용법을 논의하는 것이다. TRIPOD+AI는 예측모델 연구의 전반을 조화롭게 정비하고, 회귀분석과 머신러닝 방식 모두에 일관된 지침을 제공하는 것을 목표로 한다[66]. 여기서 "+"는 회귀분석 또는 머신러닝(딥러닝, 랜덤 포레스트 등) 접근법으로 개발된 예측모델 연구에 대해 통합된 보고 권고사항을 제공함을 의미한다. 또한 관련 연구가 일반적으로 AI로 분류되는 보고 지침과의 일관성을 위해 "AI"라는 용어가 추가되었으나, 이 논문에서는 이해를 돕기 위해 주로 머신러닝이라는 용어를 사용한다(Table 1) [67-73]. TRIPOD+AI 보고 지침에서 사용하는 주요 개념에 대한 용어 설명은 Box 1에서 확인할 수 있다.

TRIPOD+AI 개발과정

이 절에서는 머신러닝 또는 회귀방법을 이용하여 진단 또는 예후 예측모델을 개발하거나, 이들 모델의 성능을 평가(검증)하는 연구의 보고를 지원하기 위한 지침인 TRIPOD+AI 지침의 개발과정을 기술한다. '검증된 예측모델'이라는 개념은 존재하지 않으므로[76], 이 논문에서는 혼동을 피하고 용어를 통일하기 위하여 검증(validation) 대신 평가(evaluation)라는 용어를 사용하였다(Box 1). 머신러닝이 포함된 기타 생의학 연구 유형의 보고를 위한 기존 및 개발 중인 가이드라인은 Table 1에 상세히 정리하였다. TRIPOD+AI 체크리스트는 EQUATOR Network의 권고에 따라[77], 문헌고찰과 전문가 합의과정을 통해 개발되었다. G.S.C.와 K.G.M.M.의 주도로 다양한 전문성과 경험을 반영한 위원(G.S.C., K.G.M.M., R.D.R., A.L.B., J.B.R., B.V.C., X.L., P.D.)을 선정하여 지침 개발을 감독할 운영위원회를 구성하였다.

2019년 4월에는 TRIPOD+AI 이니셔티브를 알리는 해설 논문이 발표되었으며[78], 2019년 5월 7일에는 EQUATOR Network에 개발 중인 보고 지침으로 공식 등록되었다(https://www.equator-network.org/). 2021년 3월 25일에는 Open Science Framework (https://osf.io/zyacb/)에 개발과정 및 방법론을 담은 연구 프로토콜이 공개되었다. 이 프로토콜은 머신러닝 기반 예측모델의 품질 평가 및 편향 위험도구(PROBAST+AI)의 개발과정도 포함하고 있으며, 2021년에 출판되었다[79]. TRIPOD+AI 개발과정에서 활용한 합의 기반 방법의 보고는 ACCORD (Accurate Consensus Reporting Document) 권고를 준수하였다[80].

윤리적 고려

본 연구는 2020년 12월 10일 옥스퍼드대학교 중앙 연구윤리위원회(Central University Research Ethics Committee, University of Oxford)의 승인을 받았다(R73034/RE001). 델파이(Delphi) 설문 참여자에게는 설문 시작 전에, 그리고 합의 회의 참여자에게는 회의 시작 전에 참여자 안내문을 전자적으로 제공하였다. 델파이 설문 참여자는 설문 응답 전 전자 동의서를 제출하였다.

후보 항목 목록 도출

G.S.C.와 K.G.M.M.이 TRIPOD 2015 [16,17]를 바탕으로 초기 항목 목록의 초안을 작성하였다. 이후, TRIPOD-Cluster [19,20], TRIPOD for Abstracts [18], CAIR [81], MI-CLAIM [82], CLAIM [68], MINIMAR [83], SPIRIT-AI [71], CONSORT-AI [72] 및 운영위원회가 추가로 확인한 문헌[34,84-89]을 참고하여 추가 항목을 도출하였다. 머신러닝 기반 예측모델 연구의 보고, 방법론, 과도한 해석을 평가한 체계적 문헌고찰 결과[37-39,48,51,54,59,60]도 항목 목록 구성에 반영하였다. 운영위원회

Table 1. 머신러닝을 활용한 보건의료 연구의 보고 지침

보고 지침(reporting guideline)	적용 범위(scope)
STARD-AI	인공지능 기반 진단 정확도 평가 연구(작성 중)[67]
TRIPOD+AI	인공지능(머신러닝 방법 포함)을 이용한 예측 모델 개발 또는 성능 평가 연구
CLAIM	인공지능을 활용한 의료영상 연구[68]
DECIDE-AI	인공지능 기반 의사결정 지원시스템의 초기 임상 평가(안전성, 인간 요인 평가 포함)[69]
CHEERS-AI	인공지능 중재의 비용 효과성 등 건강경제학적 평가 연구[70]
SPIRIT-AI	인공지능 요소가 포함된 중재의 임상시험 연구 프로토콜[71]
CONSORT-AI	인공지능 요소가 포함된 중재의 임상시험 보고서[72]
PRISMA-AI	인공지능 중재에 관한 체계적 문헌고찰 및 메타분석(작성 중)[73]

STARD, 진단 정확도 보고 기준(Standards for Reporting of Diagnostic Accuracy); TRIPOD, 개인 예후 또는 진단을 위한 다변량 예측모델의 투명한 보고(Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis); AI, 인공지능(artificial intelligence); CLAIM, 의료영상 인공지능 연구 체크리스트(Checklist for Artificial Intelligence in Medical Imaging); DECIDE, 근거 기반 혁신의 도입 및 확산을 위한 보건의료 의사결정(Decisions in health Care to Introduce or Diffuse innovations using Evidence); CHEERS, 건강경제학적 평가 통합 보고 기준(Consolidated Health Economic Evaluation Reporting Standards); SPIRIT, 중재 임상시험 프로토콜 권고(Standard Protocol Items: Recommendations for Interventional Trials); CONSORT, 임상시험 보고 통합 기준(Consolidated Standards of Reporting Trials); PRISMA, 체계적 문헌고찰 및 메타분석 보고 권고(Preferred Reporting Items for Systematic Reviews and Meta-Analyses).

Box 1. TRIPOD+AI에서 사용된 용어 해설

아래 정의 및 설명은 TRIPOD+AI* 가이드라인의 맥락에 한정된 것이며, 다른 연구 분야에는 반드시 적용되지 않을 수 있다.

인공지능(artificial intelligence): 통상적으로 인간의 지능이 필요한 과업을 수행할 수 있는 모델 및 알고리즘을 개발하는 컴퓨터 과학 분야.

보정(calibration): 관찰된 결과와 모델에서 추정된 값 간의 일치 정도. 보정은 일반적으로 추정값(x축)과 관찰값(y축)을 그래프로 나타내고, 개별 데이터의 유연한 보정 곡선을 함께 제시하여 평가하는 것이 가장 바람직하다.

진료 경로(care pathway): 특정 건강 문제 관리 또는 환자의 진료 전 과정을 포괄하는 구조적 · 조정된 진료계획.

클래스 불균형(class imbalance): 결과 사건이 발생한 집단과 발생하지 않은 집단의 빈도가 불균등한 현상.

변별력(discrimination): 모델의 예측이 결과 발생 집단과 미발생 집단을 얼마나 잘 구분하는지의 정도. 변별력은 이항 결과의 경우 c-통계량(또는 곡선하면적[area under the curve], 수신자조작특성곡선하영역[area under the receiver operating characteristic curve])으로, 시점-사건(time-to-event) 결과는 c-지수로 정량화된다.

평가 또는 테스트 데이터(evaluation or test data): 예측모델의 성능을 추정하는 데 사용되는 데이터. ‘테스트 데이터’ 또는 ‘검증 데이터’로도 불린다.[a] 평가 데이터는 모델 훈련, 하이퍼파라미터 튜닝, 모델 선택 등에 사용된 데이터와 구분되어야 하며, 두 데이터 세트 간 참가자의 중복이 없어야 한다. 평가 데이터는 모델이 실제로 사용될 대상 인구를 대표해야 한다.

공정성(fairness): 예측모델이 연령, 인종/민족, 성별/젠더, 사회경제적 지위 등과 같은 특성을 바탕으로 개인 또는 집단을 차별하지 않는 특성.

하이퍼파라미터(hyperparameters): 모델 개발 또는 학습과정을 제어하는 값.

하이퍼파라미터 튜닝(hyperparameter tuning): 특정 모델 구축 전략에 가장 적합한 (하이퍼)파라미터 설정을 찾는 과정.

내부 검증(internal validation): 모델이 개발된 동일한 집단을 대상으로 예측모델의 성능을 평가하는 것(예: 훈련-테스트 분할, 교차검증, 부트스트래핑[bottstrapping] 등).

머신러닝(machine learning): 데이터로부터 명시적으로 프로그래밍하지 않고 학습하고 예측이나 의사결정을 내릴 수 있는 모델을 개발하는 인공지능의 한 분야.

모델 평가(model evaluation): c-통계량 등으로 모델의 변별력, 보정도(보정도 그래프, 보정 기울기 등), 임상적 유용성(의사결정 곡선 분석 등)을 추정하여 모델의 예측 정확도를 평가하는 과정. 이 과정을 예측모델의 평가라 부른다[74,75].

결과(outcome): 예측하고자 하는 진단 또는 예후 사건. 머신러닝에서는 이를 목표값(target value), 반응변수(response variable), 또는 레이블(label)이라고 지칭하기도 한다.

예측 변수(predictor): 개인 수준(예: 나이, 수축기 혈압, 성별, 질병 단계, 라디오믹스 특성) 또는 집단 수준(예: 국가)에서 측정되거나 할당될 수 있는 특성. 입력값, 특성(feature), 독립변수, 공변량 등으로도 불린다.

훈련 또는 개발 데이터(training or development data): 예측모델의 훈련 또는 개발에 사용되는 데이터. 이상적으로는, 훈련 데이터가 모델 실제 사용 인구를 대표해야 한다.

TRIPOD, 개인별 예후 또는 진단을 위한 다변량 예측모델의 투명한 보고(transparent reporting of a multivariable prediction model for individual prognosis or diagnosis); AI, 인공지능(artificial intelligence).

[a] 검증 데이터(validation data)는 연구마다 의미가 다를 수 있다. 예를 들어, 머신러닝 연구에서 검증 데이터는 파라미터 튜닝에 사용되는 데이터 또는 모델 성능 평가(대개 외부 검증이라고도 함)에 사용되는 데이터를 의미할 수 있다. 이 가이드라인에서는 혼동을 방지하기 위해 모델 성능 평가에 사용된 데이터를 평가 데이터(evaluation data)라 명명하였다.

는 이러한 자료를 바탕으로, 제목(1항목), 초록(1항목), 서론(3항목), 방법(37항목), 결과(15항목), 논의(5항목), 기타(3항목)를 포함하는 65개의 고유 후보 항목으로 최종 목록을 정비하였다. 이 목록은 이후 설명하는 수정 델파이 합의과정에서 활용되었다.

델파이 패널 선정

델파이 조사 참여자는 운영위원회가 선정하였으며, 관련 논문 저자, 소셜미디어(예: 트위터) 모집공고, 그리고 개인 추천을 통해 모집하였다. 이에는 다른 델파이 참여자가 추천한 전문가도 포함된다. 운영위원회는 지리적 및 학문적 다양성을 확보하고, 주요 이해관계자 집단—예를 들어 연구자(통계학자, 데이터 과학자, 역학자, 머신러닝 연구자, 임상의, 영상의학과 전문의, 윤리학자 등), 의료 전문가, 학술지 편집자, 연구비 지원기관, 정책 입안자, 보건의료 규제기관, 예측모델의 실제 사용자(환자 및 일반 대중 등)—을 포괄하도록 참여자를 선정하였다. 참여자는 대학, 병원, 1차 의료기관, 생의학 학술지, 비영리기관, 영리기관 등 다양한 환경에서 모집하였다.

델파이 참여자의 최소 표본 수에는 제한을 두지 않았다. 선정된 모든 참여자에 대해 운영위원회 구성원이 전문성이나 관련 경험을 확인하였다. 이후 각 참여자에게 이메일로 연구설명, 목표, 연락처 등이 포함된 안내자료와 함께 참여 초대장을 발송하였다. 참여자가 동의하면 델파이 패널로 등록되어 설문 링크를 받았다. 델파이 패널은 설문 참여에 대한 금전적 보상이나 선물을 제공받지 않았다.

델파이 과정

델파이 설문조사는 Welphi 온라인 플랫폼(www.welphi.com)을 통해, 각 참여자가 개별적으로 온라인(영어)으로 응답할 수 있도록 설계되어 배포되었다. 이 플랫폼은 각 참여자에게 개별 링크를 발

송하고 응답자에게 코드를 부여하여 익명성을 보장한다. 패널에게는 연구의 목적과 범위, 참여방법, 플랫폼 사용법, 문의처 등을 포함한 안내자료를 제공하였다. 참여자들은 각 항목에 대해 '제외 가능,' '포함 가능,' '포함 권장,' '포함 필수' 중 하나로 평가하도록 요청받았다. 또한 각 항목에 대해 자유롭게 의견을 남기거나 신규 항목을 제안할 수 있었다. 자유 서술식 응답은 P.L.이 취합 및 분석하였으며, 이를 바탕으로 G.S.C.와 K.G.M.M.이 항목의 재서술, 통합, 신규 항목 제안을 논의하였다. 운영위원회 구성원 전원에게 델파이 설문 참여 기회가 주어졌다.

1차 라운드 참여자

292명에게 초대장과 설문 참여 링크가 발송되었으며, 1차 라운드는 2021년 4월 19일부터 5월 13일까지 진행되었다. 2021년 5월 5일에 확인 메시지가 발송되었다. 초대된 292명 중 170명(부분 응답자 8명 포함)이 설문을 완료하였다. 참여자는 총 22개국에서 모집되었으며, 주요 국가는 영국(n = 52), 미국(n = 31), 네덜란드(n = 23), 캐나다(n = 20)였다. 5개 대륙(유럽 100명, 남미 2명, 북미 51명, 오세아니아 4명, 아시아 13명)에서 응답하였고, 7명은 국가를 밝히지 않았다.

참여자들은 자신의 주요 연구/업무 분야를 복수 선택할 수 있었다. 통계·데이터 과학(n = 70), AI 또는 머신러닝(n = 69), 임상(n = 50), 역학(n = 40), 예측(n = 18), 영상의학(n = 18), 보건 정책/규제(n = 10), 생의학 연구(n = 7), 학술지 편집자(n = 6), 메타연구/보고(n = 6), 병리학(n = 2), 연구비 지원 기관(n = 2), 윤리(n = 2), 기술개발/실행(n = 2), 유전학/유전체(n = 2), 의생명공학(n = 2), 보건 경제(n = 2) 등이 보고되었다.

2차 라운드 참여자

2차 델파이 라운드는 2021년 12월 16일부터 2022년 1월 17일까지 진행되었다. 1차 라운드 설문을 완료한 모든 참여자가 2차 라운드에 초대되었으며, 1차 미응답자와 1차 이후 추가 추천된 참여자도 재초대되었다. 2차 라운드 초대장은 총 395명에게 발송되었고, 200명(부분 응답자 15명 포함)이 설문을 완료하였다. 응답자는 27개국에서 모집되었으며, 역시 영국(n = 70), 미국(n = 37), 네덜란드(n = 19), 캐나다(n = 19)가 다수를 차지하였다. 6개 대륙(유럽 123명, 남미 3명, 북미 56명, 오세아니아 7명, 아시아 10명, 아프리카 1명)에서 응답이 있었다. 주요 분야는 통계·데이터 과학(n = 78), AI 또는 머신러닝(n = 72), 임상(n = 49), 역학(n = 51), 예측(n = 19), 영상의학(n = 26), 보건정책(n = 12), 생의학 연구(n = 14), 학술지 편집자(n = 13), 메타연구/보고(n = 6), 의생명공학(n = 5), 연구비 지원기관(n = 2), 유전학/유전체(n = 4), 환자 대표/참여(n = 3), 보건 경제(n = 2), 윤리(n = 1) 등이었다.

체크리스트 항목의 진화(1차→2차 라운드)

수정된 델파이 1차 라운드에서는, 참여자들이 문헌고찰 및 기존 보고 체크리스트에서 도출한 65개 후보 항목을 평가하였다. 항목 포함에 대해 '포함 권장' 또는 '포함 필수'로 응답한 경우 합의한 것으로 간주하였다. 프로토콜에서 정의한 대로[79], 70% 이상 합의에 도달한 항목만 2차 라운드로 이월되었다. 70% 미만의 항목은 제외하거나 통합 또는 재서술되어 재평가 대상으로 제시되었다. 이러한 수정은 수백 건의 패널 의견을 반영하여 이루어졌다.

2차 라운드에서는 1차 라운드의 집계결과(https://osf.io/zyacb/)를 참고하도록 안내하고, 59개 후보 항목(제목 1, 초록 1, 서론 4, 방법 32, 결과 11, 논의 8, 기타 2)에 대해 평가하도록 하였다. 환자 및 공공 참여 관련 항목은 포함 합의율이 69%로 70% 기준에 약간 못 미쳤으나, 운영위원회는 합의 회의에서 이 항목을 논의 항목으로 유지하기로 결정하였다.

환자 및 대중 참여 회의

2022년 4월 8일, Health Data Research UK의 환자 및 대중 참여 그룹(Patient and Public Involvement and Engagement, PPIE; https://www.hdruk.ac.uk/about-us/involving-and-engaging-patients-and-the-public/) 소속 9명을 대상으로 온라인 회의가 진행되었다. 이 회의는 University of Warwick의 Sophie Staniszewska가 주재하였다. 이 회의는 연구 프로토콜에 계획되어 있지 않았으며, 출판된 프로토콜[79]과의 유일한 차이점이었다. PPIE 그룹은 회의에 앞서, TRIPOD+AI 프로젝트의 요약(https://osf.io/zyacb/ 참조), PPIE 그룹원 한 명이 작성한 요약문, 그리고 체크리스트 초안을 전달받았다. 회의에서 GSC는 TRIPOD+AI 이니셔티브의 세부 내용, 프로젝트 현황, 2차 델파이 설문결과를 바탕으로 한 초안 지침을 발표하였다. 이후 참여자들은 질의응답을 진행하며 프로젝트의 목표와 범위에 대해 논의하였다. 명확성을 높이기 위해, 회의 중 제기된 의견과 회의 이후 받은 서면 피드백을 바탕으로 체크리스트 초안이 수정되었다. PPI 그룹의 세 명이 다양한 이해관계자가 참여한 2022년 7월 5일의 온라인 합의 회의에 초대되었으며, 이 중 두 명이 실제로 참석하였다. 원고는 세 명의 PPI 회원에게 전달되어 의견을 받고 승인절차를 거쳤다.

합의 회의(consensus meeting)

2022년 7월 5일, G.S.C.와 K.G.M.M.의 사회로 온라인 합의 회의가 개최되었다. 주요 이해관계자 그룹과 다양한 학문 분야, 지리적 다양성이 균형 있게 반영되도록 참가자를 선정하였다. 총 28명이 회의의 전부 또는 일부에 참석하였으며, 이 중 1명(P.L.)은 투표권이 없는 참관자였다. 초청받은 참가자들에게는 TRIPOD+AI 개요, 합의 회의 진행방식 및 안내, 2차 델파이 설문 종합결과 요약, 그리고 TRIPOD+AI 체크리스트 초안이 포함된 문서(https://osf.io/zyacb/)를 회의 전에 이메일로 발송하였다. 체크리스트 초

안은 제목(1항목), 초록(1항목), 서론(4항목), 방법(32항목), 결과(11항목), 논의(8항목), 기타(2항목) 등 총 59개 항목을 포함하였다.

2차 라운드에서 다수 항목에 대한 강한 지지가 확인되었기 때문에, 이 중 17개 항목이 전체 회의에서 토론 및 표결 대상이 되었다. 논의 후 각 항목에 대해 TRIPOD+AI 체크리스트 포함 여부를 1분간 투표하도록 하였으며, 온라인 회의 플랫폼의 투표 기능이 사용되었다. 이 17개 항목에는 2차 라운드에서 합의에 도달하지 못한 1개 항목과, 2차 라운드 이후 재서술되었거나 TRIPOD 2015에 포함되지 않았던 신규 항목 16개가 포함되었다. 이들 17개 항목에 대한 논의와 표결을 거쳐 최종 TRIPOD+AI 체크리스트를 확정하였다.

TRIPOD+AI 지침

TRIPOD+AI는 통계적 또는 머신러닝 방법을 이용해 예측모델을 개발하거나 평가(검증)하는 연구의 적절한 보고에 필수적인 항목들로 구성된 체크리스트로(Table 2), TRIPOD 2015의 주요 변화와 추가 사항은 Box 2에 요약되어 있다. TRIPOD+AI 체크리스트는 제목(항목 1), 초록(항목 2), 서론(항목 3, 4), 방법(항목

Table 2. 예측모델 연구 보고를 위한 TRIPOD+AI 체크리스트

섹션/주제	하부 주제	항목	개발/평가[a]	체크리스트 항목
제목	제목	1	D;E	연구가 다변량 예측모델의 개발 또는 성능 평가임을, 대상 집단 및 예측할 결과와 함께 명시한다.
초록	초록	2	D;E	TRIPOD+AI 초록 체크리스트 참조
서론	배경	3a	D;E	보건의료 맥락(진단 또는 예후 등) 및 예측모델 개발/평가의 근거를 설명하고, 기존 모델에 대한 참고문헌을 포함한다.
		3b	D;E	대상 집단과 예측모델의 진료 경로 내 의도된 목적 및 사용자를 기술한다(예: 의료인, 환자, 일반인 등).
		3c	D;E	사회인구학적 집단 간 알려진 건강불평등을 기술한다.
	목적	4	D;E	연구의 목적을 구체적으로 명시하며, 예측모델의 개발 또는 검증 중 어떤 연구인지(또는 둘 다인지) 기술한다.
방법	데이터	5a	D;E	개발 및 평가 데이터의 출처를 각각 기술하고(예: 무작위 임상시험, 코호트, 진료정보, 레지스트리 등), 데이터 활용의 근거와 대표성을 설명한다.
		5b	D;E	참가자 데이터의 수집 기간(시작 및 종료), 그리고 해당 시기 종료 여부(추적 종료 등)를 명확히 한다.
	참가자	6a	D;E	연구환경의 주요 요소(예: 1차 진료, 2차 진료, 일반 인구), 기관 수와 위치를 명시한다.
		6b	D;E	연구 참가자의 선정기준을 기술한다.
		6c	D;E	적용된 치료(있는 경우)와 개발/평가과정에서의 처리방법을 설명한다.
	데이터 준비	7	D;E	데이터 전처리 및 품질 관리방법, 그리고 이 과정이 사회인구학적 집단 간 유사했는지 여부를 설명한다.
	결과	8a	D;E	예측하는 결과 및 평가 시점, 결과 선정의 근거, 결과 평가방법이 사회인구학적 집단에서 일관되게 적용됐는지 명확히 기술한다.
		8b	D;E	결과 평가에 주관적 해석이 필요한 경우, 평가자의 자격 및 인구통계적 특성을 설명한다.
		8c	D;E	예측결과 평가의 눈가림 수행 여부 및 방법을 보고한다.
	예측변수	9a	D	초기 예측변수의 선정 근거(문헌, 기존 모델, 가용 변수 등) 및 모델 구축 전 사전 선정과정을 설명한다.
		9b	D;E	모든 예측변수를 명확히 정의하고, 측정 시점과 방법(및 결과/다른 예측변수의 눈가림 여부 포함)을 기술한다.
		9c	D;E	예측변수의 측정에 주관적 해석이 필요한 경우, 평가자의 자격 및 인구통계적 특성을 설명한다.
	표본크기	10	D;E	연구 규모 산출근거를(개발/평가별로) 설명하고, 연구질문에 충분한 규모였음을 정당화하며, 표본크기 산출 세부 내용을 포함한다.
	결측 데이터	11	D;E	결측 데이터 처리 방법 및 누락 사유를 기술한다.
	분석방법	12a	D	데이터 사용(개발/성능 평가 목적 등) 및 분석방법, 데이터 분할 여부와 표본크기 요건 고려사항을 명시한다.
		12b	D	모델 유형에 따라 예측변수의 분석 처리(함수형, 재조정, 변환, 표준화 등)를 설명한다.
		12c	D	모델 유형, 근거[b], 모든 모델 구축 단계(하이퍼파라미터 튜닝 등), 내부 검증방법을 명시한다.
		12d	D;E	집단 간(병원, 국가 등) 모델 파라미터 및 성능 추정치의 이질성 처리 및 정량화 방법을 기술한다. 추가사항은 TRIPOD-Cluster 참조.[c]

(Continued on the next page)

Table 2. Continued

섹션/주제	하부 주제	항목	개발/평가[a]	체크리스트 항목
		12e	D;E	모델 성능 평가에 사용된 모든 지표 및 그래프(근거 포함)를 명시하고, 필요한 경우 여러 모델 간 비교방법도 기술한다.
		12f	E	모델 평가에서 파생된 모델 수정(재보정 등)을 전체 또는 특정 집단/환경별로 기술한다.
		12g	E	모델 평가 시, 예측값 산출방식(수식, 코드, 오브젝트, API 등)을 설명한다.
	클래스 불균형	13	D;E	클래스 불균형 처리방법, 적용 이유, 사후 재보정 방법을 기술한다.
	공정성	14	D;E	모델 공정성 향상을 위한 방법 및 근거를 설명한다.
	모델 산출값	15	D	예측모델의 산출값(확률, 분류 등)을 명확히 하고, 분류기준 및 임계값 선정방법을 상세히 설명한다.
	개발-평가 차이	16	D;E	개발 데이터와 평가 데이터 간 환경, 선정기준, 결과, 예측변수의 차이를 기술한다.
	윤리 승인	17	D;E	연구를 승인한 기관윤리위원회 또는 윤리위원회의 명칭과, 연구 참가자의 동의(또는 윤리위원회의 동의 면제) 절차를 명시한다.
오픈 사이언스	연구비	18a	D;E	본 연구의 연구비 출처 및 후원자 역할을 기술한다.
	이해관계	18b	D;E	모든 저자의 이해관계 및 재정적 공시를 명시한다.
	프로토콜	18c	D;E	연구 프로토콜의 접근 가능 위치를 알리고, 프로토콜 미작성 시에는 해당 사실을 명시한다.
	등록	18d	D;E	연구 등록정보(등록기관, 등록번호 포함)를 제공하고, 미등록 시에는 해당 사실을 명시한다.
	데이터 공유	18e	D;E	연구 데이터의 접근 가능성 및 공유방식을 기술한다.
	코드 공유	18f	D;E	분석코드의 접근 가능성 및 공유방식을 기술한다.[d]
환자 및 공공 참여	환자 및 공공 참여	19	D;E	연구설계, 수행, 보고, 해석, 확산 중 어느 단계에서든 환자/공공 참여 내역을 상세히 기술하거나, 참여가 없음을 명시한다.
결과	참가자	20a	D;E	연구 내 참가자 흐름(결과 발생 유무별 참가자 수, 추적관찰 요약 포함)을 기술하고, 필요 시 도식화한다.
		20b	D;E	전체 및 환경별 주요 특성(날짜, 주요 예측변수, 치료내역, 표본 수, 결과 발생 수, 추적기간, 결측 데이터 등)을 보고하고, 인구집단별 차이도 명시한다.
		20c	E	모델 평가에서 주요 예측변수(인구통계, 예측변수, 결과 등)의 개발 데이터와의 분포 비교를 제시한다.
	모델 개발	21	D;E	각 분석(모델 개발, 하이퍼파라미터 튜닝, 평가 등)별 참가자 수 및 결과 사건 수를 명시한다.
	모델 명세	22	D	예측모델(수식, 코드, 오브젝트, API 등) 상세 내역을 제공하고, 새로운 개인 예측 또는 제3자 평가 · 구현에 필요한 접근 제한 여부(무료, 독점 등)를 명확히 기술한다.[e]
	모델 성능	23a	D;E	신뢰구간을 포함한 모델 성능 추정치, 주요 하위집단(예: 사회인구학적)별 성능, 시각화 자료(그래프 등) 제시를 고려한다.
		23b	D;E	집단 간 모델 성능의 이질성이 평가된 경우 결과를 보고한다. 추가 내용은 TRIPOD-Cluster 참고[c]
	모델 수정	24	E	모델 수정(예: 업데이트, 재보정) 및 수정 후 성능 결과를 보고한다.
논의	해석	25	D;E	주요 결과에 대한 종합적 해석을 제시하고, 목적 및 기존 연구 맥락에서 공정성 문제를 논의한다.
	한계	26	D;E	비대표성 표본, 표본크기, 과적합, 결측 데이터 등 연구의 한계 및 이로 인한 편향, 통계적 불확실성, 일반화 가능성에 미치는 영향을 논의한다.
	활용성	27a	D	입력 데이터(예측변수 등) 품질이 낮거나 제공 불가할 때의 평가 및 처리방식을 설명한다.
		27b	D	모델 적용 및 입력 데이터 활용 시 사용자의 상호작용 필요성, 요구되는 전문성 수준을 명확히 한다.
		27c	D;E	모델의 적용성과 일반화 가능성에 초점을 두고, 향후 연구과제를 논의한다.

TRIPOD, 개인별 예후 또는 진단을 위한 다변량 예측모델의 투명한 보고(Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis); AI, 인공지능(artificial intelligence).

[a] D: 예측모델 개발에만 해당, E: 예측모델 평가에만 해당, D;E: 개발과 평가 모두에 해당. [b] 모든 모델 구축 접근법에 대해 별도로 기술. [c] TRIPOD-Cluster는 클러스터(예: 병원, 센터 등)를 명시적으로 고려하거나 성능 이질성을 탐색하는 연구 보고 체크리스트. [d] 데이터 정제, 특성 엔지니어링, 모델 구축 및 평가 등 분석코드에 해당. [e] 신규 예측 위험 추정을 위한 모델 구현 코드에 해당.

Box 2. TRIPOD 2015의 주요 변경 및 추가 사항

- **새로운 체크리스트:** 랜덤 포레스트, 딥러닝 등 어떠한 회귀 또는 머신러닝 방법을 사용한 예측모델 연구도 포함할 수 있도록 보고 권고사항을 새롭게 마련하였고, 회귀 및 머신러닝 커뮤니티 간 용어를 통합하였음.
- **TRIPOD+AI 체크리스트 도입:** TRIPOD+AI 체크리스트가 기존 TRIPOD 2015 체크리스트를 대체하므로, 더 이상 TRIPOD 2015는 사용하지 않아야 함.
- **공정성에 대한 강조:** 공정성(Box 1)을 특별히 강조하여, 보고서에서 공정성 문제를 다루기 위해 어떤 방법이 사용되었는지 반드시 언급하도록 하였고, 체크리스트 전반에 공정성 요소를 포함함.
- **초록 보고 지침 추가:** 초록 작성 시 참고할 수 있도록 TRIPOD+AI for Abstracts를 별도 포함함.
- **모델 성능 항목 수정:** 저자가 주요 하위집단(예: 사회인구학적 집단)에서 모델 성능을 평가할 것을 권고하도록 해당 항목을 수정함.
- **환자 및 공공 참여 항목 신설:** 연구의 설계, 수행, 보고(및 해석), 확산과정에서 환자 및 공공의 참여에 대해 상세히 기술하도록 저자에게 요청하는 항목을 새롭게 추가함.
- **오픈 사이언스 섹션 신설:** 연구 프로토콜, 등록, 데이터 공유, 코드 공유 등에 관한 하위항목을 포함한 오픈 사이언스 섹션을 도입함.

TRIPOD, 개인별 예후 또는 진단을 위한 다변량 예측모델의 투명한 보고(transparent reporting of a multivariable prediction model for individual prognosis or diagnosis); AI, 인공지능(artificial intelligence).

5–17), 오픈 사이언스 관행(항목 18), 환자 및 대중 참여(항목 19), 결과(항목 20–24), 논의(항목 25–27) 등 총 27개의 주요 항목으로 구성된다. 일부 항목은 복수의 세부 항목을 포함하고 있어, 총 52개의 체크리스트 세부 항목으로 구성된다.

TRIPOD+AI는 예측모델의 개발, 예측모델 성능 평가(검증), 또는 이 둘을 모두 다루는 연구를 포괄한다. D;E로 표기된 항목은 예측모델 개발 및 평가 연구 모두에 공통으로 적용된다(Table 2). 체크리스트 중 D로 표기된 항목은 예측모델 개발 연구에, E로 표기된 항목은 모델 성능 평가 연구에 적용된다. 예측모델의 개발과 평가를 모두 포함한 연구의 경우, 모든 체크리스트 항목이 적용된다.

TRIPOD+AI는 예측모델 연구의 학술지 또는 학회 초록을 위한 별도의 체크리스트도 포함하고 있다. 이 체크리스트는 기존 TRIPOD for Abstracts 지침을 업데이트한 것으로[18], 새로운 내용을 반영하고 TRIPOD+AI와의 일관성을 유지하도록 설계되었다(Table 3).

TRIPOD+AI의 권고사항은 예측모델 연구의 수행과정을 투명하게 보고하도록 안내하는 것이며, 예측모델 개발 또는 평가방법 자체를 규정하는 것은 아니다. 이 체크리스트는 연구의 질을 평가하는 도구가 아니다. 예측모델의 질과 편향 위험성 평가에는 PROBAST [90,91] 및 곧 공개될 PROBAST+AI [79] 사용을 권장하며, 관련 정보는 https://www.probast.org/에서 확인할 수 있다.

TRIPOD+AI 사용방법

TRIPOD+AI 체크리스트는 기존 TRIPOD 2015 체크리스트를 대체하므로, 이제 더 이상 TRIPOD 2015 체크리스트는 사용하지 않아야 한다. 만약 예측모델 연구에서 클러스터(예: 다수 병원, 다수 데이터 세트)를 고려했다면, 저자는 TRIPOD-Cluster의 추가 보고 권고사항[19,20]을 참고해야 한다. 2015년판 설명 및 해설 문서는 여전히 대부분의 TRIPOD+AI 보고 항목에 대한 배경 및 예시를 제공하는 중요한 자료로 남아 있다[17] (많은 항목이 변경되지 않았거나 최소한만 변경되었기 때문이며, TRIPOD+AI에 대한 보다 상세하고 최신의 해설 문서는 별도로 제작 중이다). TRIPOD+AI는 논문 작성 초기 단계부터 활용하여 모든 핵심 세부사항을 누락 없이 보고할 것을 권장한다. 각 항목별로 간단한 근거와 안내를 담은 목록형 확장 체크리스트(Supplement 1)를 개발하여, TRIPOD+AI의 실제 적용을 지원하고자 하였다.

TRIPOD+AI 체크리스트의 많은 항목은 논문 내에서 자연스러운 순서로 배열되지만, 일부 항목은 그렇지 않을 수 있다. 예측모델 논문이나 출판물 내에서 각 권고사항이 반드시 어디에 위치해야 하는지 구조적 형식을 별도로 규정하지 않으며, 해당 순서는 학술지의 투고양식에 따라 달라질 수 있다.

TRIPOD+AI에 담긴 권고사항은 최소한의 보고 기준이므로, 저자는 추가적인 정보를 제공할 수 있다. 논문 본문의 분량 제한이나 표·그림 개수 제한 등으로 인해 보고가 어려운 경우, 일부 요청된 정보나 추가 자료는 보충자료로 보고하고 그 위치를 본문에서 명시하면 된다. 필요한 정보가 이미 공개적으로 접근 가능한 연구 프로토콜에 보고되어 있다면, 해당 문서를 참조하는 것으로 충분하다. 특정 체크리스트 항목을 알 수 없거나 해당이 되지 않아 보고할 수 없다면 그 사실을 명확히 밝혀야 한다. 보충자료에 포함되지 않은 추가 파일이나 연구자료는 Open Science Framework, Dryad, figshare 등과 같은 범용 또는 소속기관의 오픈 액세스 저장소에 영구적으로 공개해야 한다. 추가 파일의 접근 정보(예: doi 번호)는 논문 본문이나 출판물에 반드시 명시·연결해야 한다.

저자는 각 항목이 본문 내 어디에 보고되어 있는지(페이지 또는 줄 번호) 명시한 완성된 체크리스트를 제출하도록 권장되는데, 이는 편집 및 동료 평가과정에 도움이 된다. 별도 작성용 TRIPOD+AI 체크리스트 양식은 Supplement 2 및 www.tripod-statement.org에서 다운로드할 수 있다.

TRIPOD+AI 관련 소식, 공지, 정보는 TRIPOD 웹사이트(www.tripod-statement.org)와 X (구 트위터, @TRIPODStatement) 등 소셜미디어 계정에서 확인할 수 있다. 또한 건강 연구의

Table 3. 학술지 또는 학회 초록에 포함해야 할 예측모델 연구의 필수 항목(TRIPOD+AI for Abstracts[a])

섹션 및 항목	체크리스트 항목
제목	1. 연구가 다변량 예측모델의 개발 또는 성능 평가임을, 대상 집단 및 예측할 결과와 함께 명시한다.
배경	2. 보건의료 맥락 및 모든 모델의 개발/성능 평가근거를 간략하게 설명한다.
목적	3. 연구목적을 구체적으로 명시하며, 모델 개발, 평가 또는 둘 다에 해당하는지 포함한다.
방법	4. 데이터 출처를 설명한다.
	5. 데이터 수집 시 적용된 선정기준과 환경을 설명한다.
	6. 예측모델이 예측하고자 하는 결과(예후모델의 경우 예측기간 포함)를 명시한다.
	7. 모델 유형, 모델 구축 단계 요약, 내부 검증방법[b]을 명시한다.
	8. 모델 성능 평가에 사용된 지표(예: 변별도, 보정, 임상적 유용성 등)를 명확히 기술한다.
결과	9. 참가자 수 및 결과 사건 수를 보고한다.
	10. 최종 모델의 예측변수를 요약한다†.
	11. 신뢰구간을 포함한 모델 성능 추정치를 보고한다.
고찰	12. 주요 결과에 대한 종합적 해석을 제시한다.
등록	13. 등록번호 및 등록기관(또는 저장소) 명칭을 명시한다.

TRIPOD, 개인별 예후 또는 진단을 위한 다변량 예측모델의 투명한 보고(Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis); AI, 인공지능(artificial intelligence).
[a]이 체크리스트는 2020년에 발표된 TRIPOD for Abstracts statement [17]를 기반으로 하였으며, TRIPOD+AI statement와의 일관성을 위해 개정 · 업데이트되었음.
[b]예측모델 개발 연구에만 해당되는 항목임.

질 및 투명성 향상 네트워크(EQUATOR Network; https://www.equator-network.org/)를 통해서도 TRIPOD+AI 지침이 배포 및 홍보된다. TRIPOD+AI의 다국어 번역도 적극적으로 환영하며, 번역을 희망하는 경우 교신저자에게 연락하면 된다. 번역과정은 원저자와의 협력 및 승인을 포함한 구조적이고 사전 정의된 절차를 따르도록 하며, 번역 관련 추가 안내는 TRIPOD 웹사이트에서 확인할 수 있다(www.tripod-statement.org).

고찰

TRIPOD+AI는 국제적 다기관 다학제 합의과정을 통해 개발되었다. 이 지침은 회귀분석 또는 머신러닝 방법을 활용하여 예측모델을 개발하거나 평가(검증)하는 연구에 대해 최소한의 보고 권고사항을 제공한다. 지침 개발 당시에는 최근 급속히 발전하고 있는 파운데이션 모델 및 대형 언어 모델(예: ChatGPT 등)은 별도로 고려하지 않았으므로, TRIPOD+AI는 비생성형 모델을 주요 대상으로 한다. 그러나 이 지침의 많은 원칙은 보건 분야 생성형 AI 연구의 투명성 확보에도 적용 가능하다. 앞으로 TRIPOD+AI가 계속해서 유효성을 유지하고 AI 및 머신러닝의 발전을 반영하기 위해서는, 예를 들어 생성형 접근법에 대한 명시적 반영 등 주기적인 업데이트가 필요하다.

TRIPOD+AI는 TRIPOD 2015를 개정하여 개발되었으며, 문헌의 체계적 고찰, 델파이 설문조사, 온라인 합의 회의를 기반으로 권고사항을 정립하였다. TRIPOD+AI의 보고 항목을 충실히 기술하면, 연구방법의 질적 평가, 연구결과의 투명성 향상, 과도한 해석의 방지, 재현 및 복제 가능성 제고, 예측모델의 실제 적용 등에 모두 도움이 될 것이다. 체크리스트 항목은 최소한의 보고 기준으로, 저자는 데이터, 연구설계, 방법, 분석, 결과, 논의 등에서 추가적인 세부사항을 제공하는 것이 일반적이다.

TRIPOD+AI는 TRIPOD 2015에서는 부족하거나 명확히 언급되지 않았던 공정성(fairness) 이슈를 전반에 걸쳐 강조한다[33]. 예측모델 연구에서의 공정성은 특히 의료 분야에서 매우 중요하며, AI 및 머신러닝이 임상 의사결정 지원도구로 활용되면서 더욱 주목받고 있다. 이 맥락에서의 공정성이란, 예측모델이 특정 집단에 불리하게 작용하지 않으며, 기존의 건강불평등을 심화하지 않고(이상적으로는 이를 완화 · 개선하는 방향으로) 설계 · 사용됨을 의미한다[92]. 공정성의 중요한 측면 중 하나는 모델 개발과 평가에 활용되는 데이터가 대표성과 다양성을 갖추고, 데이터 편향의 한계를 인지 · 관리 · 완화하는 것이다. 현재 STANDING Together 이니셔티브에서는 AI 보건 데이터 세트의 다양성, 포용성, 일반화 가능성을 높이기 위한 표준을 개발 중이다[62].

이상적으로는, 데이터에는 다양한 연령, 성별/젠더, 인종 · 민족, 건강상태 또는 동반 질환, 지역적 배경의 정보가 모두 포함되어야 하며, 이러한 다양성이 예측모델의 실제 사용 대상 인구를 대표해야 한다. 만약 모델 개발에 사용된 데이터가 의도한 전체 인구집단을 충분히 반영하지 못한다면, 데이터에 포함되지 않은 집단에서는 해당 모델이 기대한 만큼의 성능을 보이지 않을 수 있음을 명확히 밝혀야 한다. 모델 평가에 사용된 데이터가 목표 인구집단을 대표하지 못한다면, 특정 하위집단(개인적, 사회적, 임상적 속성별)의 예측 정확도 추정에 편향이 생기거나 오해를 일으킬 수 있다.

데이터 세트 내 소수 집단 또는 의료 소외 집단의 충분한 대표성 확보는 공정성을 달성하기 위한 핵심 요소이지만, 단순한 대표성만

Table 4. TRIPOD+AI 보고 지침 준수: 이해관계자별 잠재적 이익

사용자/이해관계자	권장 조치	잠재적 이익
학술기관	연구자에게 예측모델 개발, 평가, 적용 시 TRIPOD+AI 준수 권장 또는 의무화	예측모델 연구의 설계, 분석, 보고의 투명성 문화 증진
	초기 경력 연구자를 대상으로 투명하고 완전한 보고의 중요성과 이점을 교육, TRIPOD+AI 지침에 맞는 논문 · 학위 논문 작성 권장	산출 연구의 질, 책임성, 재현성, 복제 가능성, 유용성 향상
연구자	논문 작성 시 TRIPOD+AI 준수	보고의 완결성과 질 향상
		예측모델 논문에 요구되는 최소한의 세부 정보에 대한 인식 증가
		산출 연구의 질, 책임성, 재현성, 복제 가능성, 유용성 향상
		모델의 독립적 평가를 용이하게 하는 세부 정보 보고 증가
학술지 편집자	논문 제출 시 저자에게 TRIPOD+AI 및 체크리스트 작성 요구 또는 의무화	예측모델 논문에 대한 학술지 요구사항과 기대치에 대한 이해도 향상
	심사자에게 TRIPOD+AI 활용 권장	저자의 이해도 향상에 따른 심사 효율성 증가
		출판 논문의 질, 책임성, 재현성, 복제 가능성, 유용성 향상
심사위원	보고의 완결성 평가에 TRIPOD+AI 사용	심사 효율성과 질 향상
		누락된 중요 정보에 대한 구체적 피드백 제공 용이
연구비 지원기관	연구자가 연구비 신청 시 TRIPOD+AI 사용 권장 또는 의무화	연구결과의 활용성 증대, 불충분한 보고로 인한 연구 낭비 감소
	연구비 수혜 연구가 타인에게도 활용될 수 있도록 보장	
환자, 공공, 연구 참여자	저자, 심사자, 학술지, 연구비 지원기관의 TRIPOD+AI 준수 옹호	연구결과에 대한 신뢰도 향상
		예측모델 연구에 대한 이해도 증진, 연구 내 건강형평성 고려 촉진
		정밀의료 및 맞춤형 질환 관리에서 환자 보고 결과와 임상연구 결과 정렬
체계적 문헌고찰자/메타연구자	TRIPOD+AI로 보고 완결성 평가	위험도 평가도구와 병행 시 연구의 질 평가 향상(예: PROBAST)
	질 및 편향 평가 시 TRIPOD+AI 참고	메타분석에 필요한 데이터 확보 용이
정책 결정자	연구의 투명하고 완전한 보고를 위해 TRIPOD+AI 활용 권장 또는 의무화	예측모델 평가 또는 적용 결정이 완전하고 투명하게 보고된 정보에 근거하도록 보장
		근거 기반 정책 권고의 신뢰성 제고
규제 기관	임상 심사자가 의료기기 소프트웨어 등 예측모델 기반 제품 규제 심사 시 TRIPOD+AI로 임상시험 보고 완결성 평가	보고된 사용 목적과 규제상 의도 일치 확인
		의료기기 규제 심사 및 주요 임상시험 보고에서 모범사례와 일치 유도
		공통 표준 도입 유도로 제조사의 임상시험 보고 공개 장려
기술/의료기기 제조사	기술/기기 개발 · 제조에 필요한 모델 정보의 충분성 검증	공통 표준 도입 유도로 제조사의 임상시험 보고 공개 장려
의료인		
	구매 · 임상 활용 전 충분한 모델 정보 확인	모델 적용 대상군 및 지원 임상적 결정에 대한 이해도 향상
		예측결과에 대한 이해도와 한계 인식 증가
		연구결과에 대한 신뢰도 향상

TRIPOD, 개인별 예후 또는 진단을 위한 다변량 예측모델의 투명한 보고(Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis); AI, 인공지능(artificial intelligence).

으로는 완전한 공정성이 보장되지 않는다[61,93]. 이에 따라, TRIPOD+AI는 체크리스트 전반에 걸쳐 공정성 관련 항목을 포함하고 있다(배경 3c; 방법 5a, 7, 8a, 8b, 9c, 12f, 14; 결과 20b, 23a; 논의 25, 26번 항목 등).

의료 분야의 공정성이란, 예측모델의 개발, 평가, 실제 임상경로 내 적용 및 확산과정에서 환자, 대중, 임상의 등 다양한 이해관계자를 적극적으로 참여시키는 것 또한 포함한다[94]. 다양한 관점의 참여는, 예측모델이 모든 사람의 요구를 충족시키고 공정하게 사용되도록 설계·운영되는지 확인하는 데 기여하며, 건강형평성을 촉진한다. TRIPOD+AI는 대중 및 환자 참여에 관한 19번 항목을 통해 예측모델 연구에 환자와 대중 참여를 통합을 장려하고, 단순한 형식적 절차가 아닌 오픈 사이언스 및 참여의 원칙을 촉진하며, 더 높은 임상 및 대중 수용성을 지향한다.

TRIPOD+AI는 오픈 사이언스 관행을 강조한다[35]. 오픈 사이언스 관행은 예측모델 연구의 투명성, 재현성, 연구자 간 협력 증진에 필수적이다[95]. 연구 등록, 프로토콜·데이터·코드·예측모델의 공개 등은, 타 연구자가 결과를 검증하고 새로운 데이터에서 모델 성능과 안전성을 평가할 수 있도록 한다. 오픈 사이언스는 연구자들이 서로의 성과를 기반으로 추가 연구를 진행할 수 있게 하여, 보건의료 분야 발전의 효율성을 높인다. 이는 예측모델의 정확성, 신뢰성, 완전성을 높여 결과적으로 환자 치료에도 긍정적 영향을 줄 수 있다. 데이터가 개방적으로 공유될 경우 임상의와 연구자는 더 크고 다양한 환자 데이터를 바탕으로 모델을 개발·평가할 수 있으며[96], 이는 예측 정확도 향상 및 임상적 의사결정 개선으로 이어질 수 있다. 이에 TRIPOD+AI는 자금 출처(18a), 이해상충(18b), 프로토콜 공개(18c), 연구 등록(18d), 데이터 및 코드 공유(18e, 18f) 등 오픈 사이언스 관련 항목을 포함한다.

TRIPOD+AI의 주요 사용자 및 수혜자는 논문을 집필하는 연구자, 논문을 평가하는 학술지 편집자 및 동료 평가자, 그 밖의 이해관계자(예: 학술기관, 정책 입안자, 연구비 지원기관, 규제기관, 환자, 연구 참여자, 대중 등)로 예상된다(Table 4). 이 지침은 임상 예측모델 개발 및 검증 연구, 의학 연구 논문, 소프트웨어·도구 관련 보고 등 근거 기반 보고가 필요한 모든 분야에 적용될 수 있다.

학술지 편집인과 출판사는 TRIPOD+AI의 준수를 장려하기 위해 저자 안내문에 이를 명시하고, 논문 투고 및 심사 과정에서 그 사용을 의무화하며, 권고사항의 준수를 필수 요건으로 삼는 것이 바람직하다. 연구비 지원기관 역시, 예측모델 연구의 지원 신청 시 TRIPOD+AI 권고에 따른 보고 계획 제출을 요구함으로써 연구 낭비를 최소화하고 효율적 자원 활용을 도모할 것을 권고한다.

Additional information

[1] *Centre for Statistics in Medicine, UK EQUATOR Centre, Nuffield Department of Orthopaedics, Rheumatology, and Musculoskeletal Sciences, University of Oxford, Oxford OX3 7LD, UK*

[2] *Julius Centre for Health Sciences and Primary Care, University Medical Centre Utrecht, Utrecht University, Utrecht, Netherlands*

[3] *Institute of Applied Health Research, College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK*

[4] *National Institute for Health and Care Research (NIHR) Birmingham Biomedical Research Centre, Birmingham, UK*

[5] *Department of Epidemiology, Harvard T H Chan School of Public Health, Boston, MA, USA*

[6] *Department of Development and Regeneration, KU Leuven, Leuven, Belgium*

[7] *Department of Biomedical Data Science, Leiden University Medical Centre, Leiden, Netherlands*

[8] *Department of Electrical Engineering and Computer Science, Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA, USA*

[9] *Institute of Inflammation and Ageing, College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK*

[10] *University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK*

[11] *Department of Medical Information Processing, Biometry and Epidemiology, Ludwig-Maximilians-University of Munich, Munich, Germany*

[12] *Patient representative, Health Data Research UK patient and public involvement and engagement group*

[13] *Patient representative, University of East Anglia, Faculty of Health Sciences, Norwich Research Park, Norwich, UK*

[14] *Beth Israel Deaconess Medical Center, Boston, MA, USA*

[15] *Laboratory for Computational Physiology, Massachusetts Institute of Technology, Cambridge, MA, USA*

[16] *Department of Biostatistics, Harvard T H Chan School of Public Health, Boston, MA, USA*

[17] *Institute of Health Informatics, University College London, London, UK*

[18] *British Heart Foundation Data Science Centre, London, UK*

[19] *Department of Computing, Imperial College London, London, UK*

[20] *Northwestern University Feinberg School of Medicine, Chicago, IL, USA*

[21] *Hardian Health, Haywards Heath, UK*

[22] *Section for Clinical Biometrics, Centre for Medical Data Science, Medical University of Vienna, Vienna, Austria*

[23] *Princess Margaret Cancer Centre, University Health Network, Toronto, ON, Canada*

[24] *Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada*

[25] *Department of Computer Science, University of Toronto, Toronto, ON, Canada*

[26] *Vector Institute for Artificial Intelligence, Toronto, ON, Canada*

[27] *Department of Medicine, University of Cape Town, Cape Town, South Africa*

[28] *National Institute for Health and Care Excellence, London, UK*

[29] *The BMJ, London, UK*

[30] *Department of Neurology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA*

[31] *Department of Intelligent Medical Systems, German Cancer Research Centre, Heidelberg, Germany*

[32] *Wellcome Trust, London, UK*

[33] *Alan Turing Institute, London, UK*

[34] *Department of Bioethics, Hospital for Sick Children Toronto, ON, Canada*

[35] *Genetics and Genome Biology, SickKids Research Institute, Toronto, ON, Canada*

[36] *Australian Institute for Machine Learning, University of Adelaide, Adelaide, SA, Australia*

[37] *Medicines and Healthcare products Regulatory Agency, London, UK*

[38] *Department of Health Policy and Center for Health Policy, Stanford University, Stanford, CA, USA*

[39] *Department of Learning Health Sciences, University of Michigan Medical School, Ann Arbor, MI, USA*

[40]*Department of Epidemiology, CAPHRI Care and Public Health Research Institute, Maastricht University, Maastricht, Netherlands*

ORCID

Gary S. Collins: https://orcid.org/0000-0002-2772-2316
Karel G. M. Moons: https://orcid.org/0000-0003-2118-004X
Paula Dhiman: https://orcid.org/0000-0002-0989-0623
Richard D. Riley: https://orcid.org/0000-0001-8699-0735
Andrew L. Beam: https://orcid.org/0000-0002-6657-2787
Ben Van Calster: https://orcid.org/0000-0003-1613-7450
Marzyeh Ghassemi: https://orcid.org/0000-0001-6349-7251
Xiaoxuan Liu: https://orcid.org/0000-0002-1286-0038
Johannes B. Reitsma: https://orcid.org/0000-0003-4026-4345
Maarten van Smeden: https://orcid.org/0000-0002-5529-1541
Anne-Laure Boulesteix: https://orcid.org/0000-0002-2729-0947
Jennifer Catherine Camaradou: https://orcid.org/0000-0002-5742-2840
Leo Anthony Celi: https://orcid.org/0000-0001-6712-6626
Spiros Denaxas: https://orcid.org/0000-0001-9612-7791
Alastair K. Denniston: https://orcid.org/0000-0001-7849-0087
Ben Glocker: https://orcid.org/0000-0002-4897-9356
Robert M. Golub: https://orcid.org/0009-0000-3270-0632
Hugh Harvey: https://orcid.org/0000-0003-4528-1312
Georg Heinze: https://orcid.org/0000-0003-1147-8491
Michael M. Hoffman: https://orcid.org/0000-0002-4517-1562
André Pascal Kengne: https://orcid.org/0000-0002-5183-131X
Lena Maier-Hein: https://orcid.org/0000-0003-4910-9368
Bilal A. Mateen: https://orcid.org/0000-0003-4423-6472
Melissa D. McCradden: https://orcid.org/0000-0002-6476-2165
Lauren Oakden-Rayner: https://orcid.org/0000-0001-5471-5202
Johan Ordish: https://orcid.org/0000-0001-6911-2367
Richard Parnell: https://orcid.org/0000-0003-0044-3496
Sherri Rose: https://orcid.org/0000-0002-9076-8472
Karandeep Singh: https://orcid.org/0000-0001-8980-2330
Laure Wynants: https://orcid.org/0000-0002-3037-122X
Patricia Logullo: https://orcid.org/0000-0001-8708-7003

TRIPOD+AI working group/consensus meeting participants

Gary Collins (University of Oxford, UK), Karel Moons (UMC Utrecht, Netherlands), Johannes Reitsma (UMC Utrecht, Netherlands), Andrew Beam (Harvard School of Public Health, USA), Ben Van Calster (KU Leuven, Belgium), Paula Dhiman (University of Oxford, UK), Richard Riley (University of Birmingham, UK), Marzyeh Ghassemi (Massachusetts Institute of Technology, USA), Patricia Logullo (University of Oxford, UK), Maarten van Smeden (UMC Utrecht, Netherlands), Jennifer Catherine Camaradou (Health Data Research [HDR] UK public and patient involvement group, NHS England Accelerated Access Collaborative evaluation advisory group member, National Institute for Health and Care Excellence covid-19 expert panel), Richard Parnell (HDR UK public and patient involvement group), Elizabeth Loder (The BMJ), Robert Golub (Northwestern University Feinberg School of Medicine, USA [JAMA, at the time of the consensus meeting]), Naomi Lee (National Institute for Health and Clinical Excellence, UK; The Lancet, at the time of consensus meeting), Johan Ordish (Roche, UK; Medicine and Healthcare products Regulatory Agency, UK at the time of consensus meeting), Laure Wynants (KU Leuven, Belgium), Leo Celi (Massachusetts Institute of Technology, USA), Bilal Mateen (Wellcome Trust, UK), Alastair Denniston (University of Birmingham, UK), Karandeep Singh (University of Michigan, USA), Georg Heinze (Medical University of Vienna, Austria), Lauren Oaken-Rayner (University of Adelaide, Australia), Melissa McCradden (Hospital for Sick Children, Canada), Hugh Harvey (Hardian Health, UK), Andre Pascal Kengne (University of Cape Town, South Africa), Viknesh Sounderajah (Imperial College London, UK), Lena Maier-Hein (German Cancer Research Centre, Germany), Anne-Laure Boulesteix (University of Munich, Germany), Xiaoxuan Liu (University of Birmingham, UK), Emily Lam (HDR UK public and patient involvement group), Ben Glocker (Imperial College London, UK), Sherri Rose (Stanford University, US), Michael Hoffman (University of Toronto, Canada), and Spiros Denaxas (University College London, UK). The last seven participants in this list did not attend the virtual consensus meeting.

Authors' contributions

GSC and KGMM conceived the study and this paper and are joint first authors. GSC, PL, PD, RDR, ALB, BVC, XL, JBR, MvS, and KGMM designed the surveys carried out to inform the guideline content. PL analysed the survey results and free text comments from the surveys. GSC designed the materials for the consensus meeting with input from KGMM. All authors except SR, MMH, XL, SD, BG, and ALB attended the consensus meeting. PL took consolidated notes from the consensus meeting. GSC drafted the manuscript with input and edits from KGMM. All authors were involved in revising the article critically for important intellectual content and approved the final version of the article. GSC is the guarantor of this work. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Conflict of interest

All authors have completed the ICMJE uniform disclosure form at https://www.icmje.org/disclosure-of-interest/ and declare: support from the funding bodies listed above for the submitted work; no financial relationships with any organizations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work. GSC is a National Institute for Health and Care Research (NIHR) senior investigator, the director of the UK EQUATOR Centre, editor-in-chief of *BMC Diagnostic and Prognostic Research*, and a statistics editor for *The BMJ*. KGMM is director of Health Innovation Netherlands and editor-in-chief of *BMC Diagnostic and Prognostic Research*. RDR is an NIHR senior investigator, a statistics editor for *The BMJ*, and receives royalties from textbooks *Prognosis Research in Healthcare* and *Individual Participant Data Meta-Analysis*. AKD is an NIHR senior investigator. EWL is the head of research at *The BMJ*. BG is a part time employee of HeartFlow and Kheiron Medical Technologies and holds stock options with both as part of the standard compensation package. SR receives royalties from Springer for the textbooks *Targeted Learning: Causal Inference for Observational and Experimental Data and Targeted Learning: Causal Inference for Complex Longitudinal Studies*. JCC receives honorariums as a current lay member on the UK NICE COVID-19 expert panel and a citizen partner on the COVID-END COVID-19 Evidence Network to support decision making; was a lay member on the UK NIHR AI AWARD panel in 2020-22 and is a current lay member on the UK NHS England AAC Accelerated Access Collaborative NHS AI Laboratory Evaluation Advisory Group; is a patient fellow of the European Patients' Academy on Therapeutic Innovation and a EURORDIS rare disease alumni; reports grants from the UK National Institute for Health and Care Research, European Commission, UK Cell Gene Catapult, University College London, and University of East Anglia; reports patient speaker fees from MEDABLE, Reuters Pharma events, Patients as Partners Europe, and EIT Health Scandinavia; reports consultancy fees from Roche Global, Smith, the Future Science Group and Springer Healthcare (scientific publishing), outside of the scope of the present work; and is a strategic board member of the UK Medical Research Council IASB Advanced Pain Discovery Platform initiative, Plymouth Institute of Health, and EU project Digipredict Edge AI-deployed Digital Twins for COVID-19 Cardiovascular Disease. ALB is a paid consultant for Generate Biomedicines, Flagship Pioneering, Porter Health, FL97, Tessera, FL85; has an equity stake in Generate Biomedicines; and receives research funding support from Smith, National Heart, Lung, and Blood Institute, and National Institute of Diabetes and Digestive and Kidney Diseases. No other conflicts of interests with this specific work are declared.

Funding

This research was supported by Cancer Research UK programme grant (C49297/A27294), which supports GSC and PL; Health Data Research UK, an initiative funded by UK Research and Innovation, Department of Health and Social Care (England) and the devolved administrations, and leading medical research charities, which supports GSC; an Engineering and Physical Sciences Research Council grant for "Artificial intelligence innovation to accelerate health research" (EP/Y018516/1), which supports GSC, PD, and RDR; Netherlands Organisation for Scientific Research (which supports KGMM); and University Hospitals Leuven (COPREDICT grant), Internal Funds KU Leuven (grant C24M/20/064), and Research Foundation– Flanders (grant G097322N), which supports BVC and LW. The funders had no role in considering the study design or in the collection, analysis, interpretation of data, writing of the report, or decision to submit the article for publication.

Data availability

Aggregated Delphi survey responses are available on the Open Science Framework TRIPOD+AI repository https://osf.io/zyacb/.

Acknowledgments

We thank the TRIPOD+AI Delphi panel members for their time and valuable contribution in helping to develop TRIPOD+AI statement. Full list of Delphi participants are as follows (in alphabetical order of first name): Abhishek Gupta, Adrian Barnett, Adrian Jonas, Agathe Truchot, Aiden Doherty, Alan Fraser, Alex Fowler, Alex Garaiman, Alistair Denniston, Amin Adibi, André Carrington, Andre Esteva, Andrew Althouse, Andrew

Beam, Andrew Soltan, Ane Appelt, Anne-Laure Boulesteix, Ari Ercole, Armando Bedoya, Baptiste Vasey, Bapu Desiraju, Barbara Seeliger, Bart Geerts, Beatrice Panico, Ben Glocker, Ben Van Calster, Benjamin Fine, Benjamin Goldstein, Benjamin Gravesteijn, Benjamin Wissel, Bilal Mateen, Bjoern Holzhauer, Boris Janssen, Boyi Guo, Brooke Levis, Catey Bunce, Charles Kahn, Chris Tomlinson, Christopher Kelly, Christopher Lovejoy, Clare McGenity, Conrad Harrison, Constanza Andaur Navarro, Daan Nieboer, Dan Adler, Danial Bahudin, Daniel Stahl, Daniel Yoo, Danilo Bzdok, Darren Dahly, Darren Treanor, David Higgins, David McClernon, David Pasquier, David Taylor, Declan O'Regan, Emily Bebbington, Erik Ranschaert, Evangelos Kanoulas, Facundo Diaz, Felipe Kitamura, Flavio Clesio, Floor van Leeuwen, Frank Harrell, Frank Rademakers, Gael Varoquaux, Garrett Bullock, Gary Collins, Gary Weissman, Georg Heinze, George Fowler, George Kostopoulos, Georgios Lyratzaopoulos, Gianluca Di Tanna, Gianluca Pellino, Girish Kulkarni, Giuseppe Biondi Zoccai, Glen Martin, Gregg Gascon, Harlan Krumholz, Herdiantri Sufriyana, Hongqiu Gu, Hrvoje Bogunovic, Hui Jin, Ian Scott, Ijeoma Uchegbu, Indra Joshi, Irene Stratton, James Glasbey, Jamie Miles, Jamie Sergeant, Jan Roth, Jared Wohlgemut, Javier Carmona Sanz, Jean-Emmanuel Bibault, Jeremy Cohen, Ji Eun Park, Jie Ma, Joel Amoussou, Johan Ordish, Johannes Reitsma, John Pickering, Joie Ensor, Jose L Flores-Guerrero, Joseph LeMoine, Joshua Bridge, Josip Car, Junfeng Wang, Karel Moons, Keegan Korthauer, Kelly Reeve, Laura Ación, Laura Bonnett, Laure Wynants, Lena Maier-Hein, Leo Anthony Celi, Lief Pagalan, Ljubomir Buturovic, Lotty Hook, Luke Farrow, Maarten Van Smeden, Marianne Aznar, Mario Doria, Mark Gilthorpe, Mark Sendak, Martin Fabregate, Marzyeh, Ghassemi, Matthew Sperrin, Matthew Strother, Mattia Prosperi, Melissa McCradden, Menelaos Konstantinidis, Merel Huisman, Michael Harhay, Michael Hoffman, Miguel Angel Luque, Mohammad Mansournia, Munya Dimairo, Musa Abdulkareem, Myura Nagendran, Niels Peek, Nigam Shah, Nikolas Pontikos, Nurulamin Noor, Oilivier Groot, Pall Jonsson, Patricia Logullo, Patrick Bossuyt, Patrick Lyons, Patrick Omoumi, Paul Tiffin, Paula Dhiman, Peter Austin, Quentin Noirhomme, Rachel Kuo, Ram Bajpal, Ravi Aggarwal, Richard Riley, Richiardi Jonas, Robert Golub, Robert Platt, Rohit Singla, Roi Anteby, Rupa Sakar, Safoora Masoumi, Sara Khalid, Saskia Haitjema, Seong Park, Shravya Shetty, Spiros Denaxas, Stacey Fisher, Stephanie Hicks, Susan Shelmerdine, Tammy Clifford, Tatyana Shamliyan, Teus Kappen, Tim Leiner, Tim Liu, Tim Ramsay, Toni Martinez, Uri Shalit, Valentijn de Jong, Valentyn, Bezshapkin, Veronika Cheplygina, Victor Castro, Viknesh Sounderajah, Vineet Kamal, Vinyas Harish, Wim Weber, Wouter Amsterdam, Xioaxuan Liu, Zachary Cohen, Zakia Salod, and Zane Perkins.

We thank Sophie Staniszewska (University of Warwick, UK) for chairing the HDR UK patient and public involvement and engagement meeting, where the TRIPOD+AI study and drak (pre-consensus meeting) checklist was presented and discussed; and Jennifer de Beyer for proofreading the manuscript (University of Oxford, UK).

Supplementary materials

The online version contains supplementary material available at https://doi.org/10.12771/emj.2025.00668

Supplement 1. TRIPOD+AI Expanded Checklist (Explanation & Elaboration Light).

Supplement 2. Fillable TRIPOD+AI checklist.

References

1. van Smeden M, Reitsma JB, Riley RD, Collins GS, Moons KG. Clinical prediction models: diagnosis versus prognosis. J Clin Epidemiol 2021;132:142-145. https://doi.org/10.1016/j.jclinepi.2021.01.009
2. Nashef SA, Roques F, Sharples LD, Nilsson J, Smith C, Goldstone AR, Lockowandt U. EuroSCORE II. Eur J Cardiothorac Surg 2012;41:734-745. https://doi.org/10.1093/ejcts/ezs043
3. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, Mulvihill JJ. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst 1989;81:1879-1886. https://doi.org/10.1093/jnci/81.24.1879
4. D'Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, Kannel WB. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation 2008;117:743-753. https://doi.org/10.1161/CIRCULATIONAHA.107.699579
5. Steyerberg EW, Mushkudiani N, Perel P, Butcher I, Lu J, McHugh GS, Murray GD, Marmarou A, Roberts I, Habbema JD, Maas AI. Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics. PLoS Med 2008;5:e165. https://doi.org/10.1371/journal.pmed.0050165
6. Kanis JA, Oden A, Johnell O, Johansson H, De Laet C, Brown J, Burckhardt P, Cooper C, Christiansen C, Cummings S, Eisman

JA, Fujiwara S, Glüer C, Goltzman D, Hans D, Krieg MA, La Croix A, McCloskey E, Mellstrom D, Melton LJ, Pols H, Reeve J, Sanders K, Schott AM, Silman A, Torgerson D, van Staa T, Watts NB, Yoshimura N. The use of clinical risk factors enhances the performance of BMD in the prediction of hip and osteoporotic fractures in men and women. Osteoporos Int 2007; 18:1033-1046. https://doi.org/10.1007/s00198-007-0343-y

7. Damen JA, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, Lassale CM, Siontis GC, Chiocchia V, Roberts C, Schlussel MM, Gerry S, Black JA, Heus P, van der Schouw YT, Peelen LM, Moons KG. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ 2016;353:i2416. https://doi.org/10.1136/bmj.i2416
8. Bellou V, Belbasis L, Konstantinidis AK, Tzoulaki I, Evangelou E. Prognostic models for outcome prediction in patients with chronic obstructive pulmonary disease: systematic review and critical appraisal. BMJ 2019;367:l5358. https://doi.org/10.1136/bmj.l5358
9. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, Bonten MM, Dahly DL, Damen JA, Debray TP, de Jong VM, De Vos M, Dhiman P, Haller MC, Harhay MO, Henckaerts L, Heus P, Kammer M, Kreuzberger N, Lohmann A, Luijken K, Ma J, Martin GP, McLernon DJ, Andaur Navarro CL, Reitsma JB, Sergeant JC, Shi C, Skoetz N, Smits LJ, Snell KI, Sperrin M, Spijker R, Steyerberg EW, Takada T, Tzoulaki I, van Kuijk SM, van Bussel B, van der Horst IC, van Royen FS, Verbakel JY, Wallisch C, Wilkinson J, Wolff R, Hooft L, Moons KG, van Smeden M. Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal. BMJ 2020;369:m1328. https://doi.org/10.1136/bmj.m1328
10. Mallett S, Royston P, Dutton S, Waters R, Altman DG. Reporting methods in studies developing prognostic models in cancer: a review. BMC Med 2010;8:20. https://doi.org/10.1186/1741-7015-8-20
11. Collins GS, Mallett S, Omar O, Yu LM. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. BMC Med 2011;9:103. https://doi.org/10.1186/1741-7015-9-103
12. Altman DG, Simera I, Hoey J, Moher D, Schulz K. EQUATOR: reporting guidelines for health research. Open Med 2008; 2:e49-e50.
13. Glasziou P, Altman DG, Bossuyt P, Boutron I, Clarke M, Julious S, Michie S, Moher D, Wager E. Reducing waste from incomplete or unusable reports of biomedical research. Lancet 2014;383:267-276. https://doi.org/10.1016/S0140-6736(13)62228-X
14. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, Voysey M, Wharton R, Yu LM, Moons KG, Altman DG. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol 2014;14:40. https://doi.org/10.1186/1471-2288-14-40
15. Bouwmeester W, Zuithoff NP, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, Altman DG, Moons KG. Reporting and methods in clinical prediction research: a systematic review. PLoS Med 2012;9:1-12. https://doi.org/10.1371/journal.pmed.1001221
16. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med 2015;162:55-63. https://doi.org/10.7326/M14-0697
17. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, Vickers AJ, Ransohoff DF, Collins GS. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015;162:W1-W73. https://doi.org/10.7326/M14-0698
18. Heus P, Reitsma JB, Collins GS, Damen JA, Scholten RJ, Altman DG, Moons KG, Hooft L. Transparent reporting of multivariable prediction models in journal and conference abstracts: TRIPOD for abstracts. Ann Intern Med 2020 Jun 2 [Epub]. https://doi.org/10.7326/M20-0193
19. Debray TP, Collins GS, Riley RD, Snell KI, Van Calster B, Reitsma JB, Moons KG. Transparent reporting of multivariable prediction models developed or validated using clustered data: TRIPOD-Cluster checklist. BMJ 2023;380:e071018. https://doi.org/10.1136/bmj-2022-071018
20. Debray TP, Collins GS, Riley RD, Snell KI, Van Calster B, Reitsma JB, Moons KG. Transparent reporting of multivariable prediction models developed or validated using clustered data (TRIPOD-Cluster): explanation and elaboration. BMJ 2023; 380:e071058. https://doi.org/10.1136/bmj-2022-071058
21. Snell KI, Levis B, Damen JA, Dhiman P, Debray TP, Hooft L, Reitsma JB, Moons KG, Collins GS, Riley RD. Transparent reporting of multivariable prediction models for individual prognosis or diagnosis: checklist for systematic reviews and meta-analyses (TRIPOD-SRMA). BMJ 2023;381:e073538.

https://doi.org/10.1136/bmj-2022-073538

22. Dhiman P, Whittle R, Van Calster B, Ghassemi M, Liu X, McCradden MD, Moons KG, Riley RD, Collins GS. The TRIPOD-P reporting guideline for improving the integrity and transparency of predictive analytics in healthcare through study protocols. Nat Mach Intell 2023;5:816-817. https://doi.org/10.1038/s42256-023-00705-6
23. Riley RD, Snell KI, Ensor J, Burke DL, Harrell FE, Moons KG, Collins GS. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. Stat Med 2019;38:1276-1296. https://doi.org/10.1002/sim.7992
24. Riley RD, Snell KI, Ensor J, Burke DL, Harrell FE, Moons KG, Collins GS. Minimum sample size for developing a multivariable prediction model: Part I - Continuous outcomes. Stat Med 2019;38:1262-1275. https://doi.org/10.1002/sim.7993
25. Riley RD, Ensor J, Snell KI, Harrell FE, Martin GP, Reitsma JB, Moons KG, Collins G, van Smeden M. Calculating the sample size required for developing a clinical prediction model. BMJ 2020;368:m441. https://doi.org/10.1136/bmj.m441
26. van Smeden M, de Groot JA, Moons KG, Collins GS, Altman DG, Eijkemans MJ, Reitsma JB. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. BMC Med Res Methodol 2016;16:163. https://doi.org/10.1186/s12874-016-0267-3
27. van Smeden M, Moons KG, de Groot JA, Collins GS, Altman DG, Eijkemans MJ, Reitsma JB. Sample size for binary logistic prediction models: beyond events per variable criteria. Stat Methods Med Res 2019;28:2455-2474. https://doi.org/10.1177/0962280218784726
28. Snell KI, Archer L, Ensor J, Bonnett LJ, Debray TP, Phillips B, Collins GS, Riley RD. External validation of clinical prediction models: simulation-based sample size calculations were more reliable than rules-of-thumb. J Clin Epidemiol 2021;135:79-89. https://doi.org/10.1016/j.jclinepi.2021.02.011
29. Archer L, Snell KI, Ensor J, Hudda MT, Collins GS, Riley RD. Minimum sample size for external validation of a clinical prediction model with a continuous outcome. Stat Med 2021;40:133-146. https://doi.org/10.1002/sim.8766
30. Riley RD, Debray TP, Collins GS, Archer L, Ensor J, van Smeden M, Snell KI. Minimum sample size for external validation of a clinical prediction model with a binary outcome. Stat Med 2021;40:4230-4251. https://doi.org/10.1002/sim.9025
31. Riley RD, Collins GS, Ensor J, Archer L, Booth S, Mozumder SI, Rutherford MJ, van Smeden M, Lambert PC, Snell KI. Minimum sample size calculations for external validation of a clinical prediction model with a time-to-event outcome. Stat Med 2022;41:1280-1295. https://doi.org/10.1002/sim.9275
32. Riley RD, Snell KI, Archer L, Ensor J, Debray TP, van Calster B, van Smeden M, Collins GS. Evaluation of clinical prediction models (part 3): calculating the sample size required for an external validation study. BMJ 2024;384:e074821. https://doi.org/10.1136/bmj-2023-074821
33. Wawira Gichoya J, McCoy LG, Celi LA, Ghassemi M. Equity in essence: a call for operationalising fairness in machine learning for healthcare. BMJ Health Care Inform 2021;28:e100289. https://doi.org/10.1136/bmjhci-2020-100289
34. McDermott MB, Wang S, Marinsek N, Ranganath R, Foschini L, Ghassemi M. Reproducibility in machine learning for health research: still a ways to go. Sci Transl Med 2021;13:eabb1655. https://doi.org/10.1126/scitranslmed.abb1655
35. UNESCO. UNESCO recommendation on Open Science [Internet]. UNESCO; 2023 [cited 2025 Jul 10]. Available from: https://www.unesco.org/en/open-science/about?hub=686
36. Wessler BS, Nelson J, Park JG, McGinnes H, Gulati G, Brazil R, Van Calster B, van Klaveren D, Venema E, Steyerberg E, Paulus JK, Kent DM. External validations of cardiovascular clinical prediction models: a large-scale review of the literature. Circ Cardiovasc Qual Outcomes 2021;14:e007858. https://doi.org/10.1161/CIRCOUTCOMES.121.007858
37. Dhiman P, Ma J, Andaur Navarro CL, Speich B, Bullock G, Damen JA, Hooft L, Kirtley S, Riley RD, Van Calster B, Moons KG, Collins GS. Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review. BMC Med Res Methodol 2022;22:101. https://doi.org/10.1186/s12874-022-01577-x
38. Andaur Navarro CL, Damen JA, van Smeden M, Takada T, Nijman SW, Dhiman P, Ma J, Collins GS, Bajpai R, Riley RD, Moons KG, Hooft L. Systematic review identifies the design and methodological conduct of studies on machine learning-based prediction models. J Clin Epidemiol 2023;154:8-22. https://doi.org/10.1016/j.jclinepi.2022.11.015
39. Andaur Navarro CL, Damen JA, Takada T, Nijman SW, Dhiman P, Ma J, Collins GS, Bajpai R, Riley RD, Moons KG, Hooft L. Completeness of reporting of clinical prediction models developed using supervised machine learning: a systematic review. BMC Med Res Methodol 2022;22:12. https://doi.org/10.1186/s12874-021-01469-6

40. Rech MM, de Macedo Filho L, White AJ, Perez-Vega C, Samson SL, Chaichana KL, Olomu OU, Quinones-Hinojosa A, Almeida JP. Machine learning models to forecast outcomes of pituitary surgery: a systematic review in quality of reporting and current evidence. Brain Sci 2023;13:495. https://doi.org/10.3390/brainsci13030495
41. Munguia-Realpozo P, Etchegaray-Morales I, Mendoza-Pinto C, Mendez-Martinez S, Osorio-Pena AD, Ayon-Aguilar J, Garcia-Carrasco M. Current state and completeness of reporting clinical prediction models using machine learning in systemic lupus erythematosus: a systematic review. Autoimmun Rev 2023;22:103294. https://doi.org/10.1016/j.autrev.2023.103294
42. Kee OT, Harun H, Mustafa N, Abdul Murad NA, Chin SF, Jaafar R, Abdullah N. Cardiovascular complications in a diabetes prediction model using machine learning: a systematic review. Cardiovasc Diabetol 2023;22:13. https://doi.org/10.1186/s12933-023-01741-7
43. Song Z, Yang Z, Hou M, Shi X. Machine learning in predicting cardiac surgery-associated acute kidney injury: a systemic review and meta-analysis. Front Cardiovasc Med 2022;9:951881. https://doi.org/10.3389/fcvm.2022.951881
44. Yang Q, Fan X, Cao X, Hao W, Lu J, Wei J, Tian J, Yin M, Ge L. Reporting and risk of bias of prediction models based on machine learning methods in preterm birth: a systematic review. Acta Obstet Gynecol Scand 2023;102:7-14. https://doi.org/10.1111/aogs.14475
45. Groot OQ, Ogink PT, Lans A, Twining PK, Kapoor ND, DiGiovanni W, Bindels BJ, Bongers ME, Oosterhoff JH, Karhade AV, Oner FC, Verlaan JJ, Schwab JH. Machine learning prediction models in orthopedic surgery: a systematic review in transparent reporting. J Orthop Res 2022;40:475-483. https://doi.org/10.1002/jor.25036
46. Lans A, Kanbier LN, Bernstein DN, Groot OQ, Ogink PT, Tobert DG, Verlaan JJ, Schwab JH. Social determinants of health in prognostic machine learning models for orthopaedic outcomes: a systematic review. J Eval Clin Pract 2023;29:292-299. https://doi.org/10.1111/jep.13765
47. Li B, Feridooni T, Cuen-Ojeda C, Kishibe T, de Mestral C, Mamdani M, Al-Omran M. Machine learning in vascular surgery: a systematic review and critical appraisal. NPJ Digit Med 2022;5:7. https://doi.org/10.1038/s41746-021-00552-y
48. Groot OQ, Bindels BJ, Ogink PT, Kapoor ND, Twining PK, Collins AK, Bongers ME, Lans A, Oosterhoff JH, Karhade AV, Verlaan JJ, Schwab JH. Availability and reporting quality of external validations of machine-learning prediction models with orthopedic surgical outcomes: a systematic review. Acta Orthop 2021;92:385-393. https://doi.org/10.1080/17453674.2021.1910448
49. Andaur Navarro CL, Damen JA, Takada T, Nijman SW, Dhiman P, Ma J, Collins GS, Bajpai R, Riley RD, Moons KG, Hooft L. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. BMJ 2021;375:n2281. https://doi.org/10.1136/bmj.n2281
50. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019;110:12-22. https://doi.org/10.1016/j.jclinepi.2019.02.004
51. Yusuf M, Atal I, Li J, Smith P, Ravaud P, Fergie M, Callaghan M, Selfe J. Reporting quality of studies using machine learning models for medical diagnosis: a systematic review. BMJ Open 2020;10:e034568. https://doi.org/10.1136/bmjopen-2019-034568
52. Wang W, Kiik M, Peek N, Curcin V, Marshall IJ, Rudd AG, Wang Y, Douiri A, Wolfe CD, Bray B. A systematic review of machine learning models for predicting outcomes of stroke with structured data. PLoS One 2020;15:e0234722. https://doi.org/10.1371/journal.pone.0234722
53. Miles J, Turner J, Jacques R, Williams J, Mason S. Using machine-learning risk prediction models to triage the acuity of undifferentiated patients entering the emergency care system: a systematic review. Diagn Progn Res 2020;4:16. https://doi.org/10.1186/s41512-020-00084-1
54. Dhiman P, Ma J, Navarro CA, Speich B, Bullock G, Damen JA, Kirtley S, Hooft L, Riley RD, Van Calster B, Moons KG, Collins GS. Reporting of prognostic clinical prediction models based on machine learning methods in oncology needs to be improved. J Clin Epidemiol 2021;138:60-72. https://doi.org/10.1016/j.jclinepi.2021.06.024
55. Dhiman P, Ma J, Andaur Navarro CL, Speich B, Bullock G, Damen JA, Hooft L, Kirtley S, Riley RD, Van Calster B, Moons KG, Collins GS. Risk of bias of prognostic models developed using machine learning: a systematic review in oncology. Diagn Progn Res 2022;6:13. https://doi.org/10.1186/s41512-022-00126-w
56. Araujo AL, Moraes MC, Perez-de-Oliveira ME, Silva VM, Saldivia-Siracusa C, Pedroso CM, Lopes MA, Vargas PA, Ko-

channy S, Pearson A, Khurram SA, Kowalski LP, Migliorati CA, Santos-Silva AR. Machine learning for the prediction of toxicities from head and neck cancer treatment: a systematic review with meta-analysis. Oral Oncol 2023;140:106386. https://doi.org/10.1016/j.oraloncology.2023.106386
57. Sheehy J, Rutledge H, Acharya UR, Loh HW, Gururajan R, Tao X, Zhou X, Li Y, Gurney T, Kondalsamy-Chennakesavan S. Gynecological cancer prognosis using machine learning techniques: a systematic review of the last three decades (1990-2022). Artif Intell Med 2023;139:102536. https://doi.org/10.1016/j.artmed.2023.102536
58. Collins GS, Whittle R, Bullock GS, Logullo P, Dhiman P, de Beyer JA, Riley RD, Schlussel MM. Open science practices need substantial improvement in prognostic model studies in oncology using machine learning. J Clin Epidemiol 2024;165:111199. https://doi.org/10.1016/j.jclinepi.2023.10.015
59. Dhiman P, Ma J, Andaur Navarro CL, Speich B, Bullock G, Damen JA, Hooft L, Kirtley S, Riley RD, Van Calster B, Moons KG, Collins GS. Overinterpretation of findings in machine learning prediction model studies in oncology: a systematic review. J Clin Epidemiol 2023;157:120-133. https://doi.org/10.1016/j.jclinepi.2023.03.012
60. Andaur Navarro CL, Damen JA, Takada T, Nijman SW, Dhiman P, Ma J, Collins GS, Bajpai R, Riley RD, Moons KG, Hooft L. Systematic review finds “spin” practices and poor reporting standards in studies on machine learning-based prediction models. J Clin Epidemiol 2023;158:99-110. https://doi.org/10.1016/j.jclinepi.2023.03.024
61. Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical machine learning in healthcare. Annu Rev Biomed Data Sci 2021;4:123-144. https://doi.org/10.1146/annurev-biodatasci-092820-114757
62. Ganapathi S, Palmer J, Alderman JE, Calvert M, Espinoza C, Gath J, Ghassemi M, Heller K, Mckay F, Karthikesalingam A, Kuku S, Mackintosh M, Manohar S, Mateen BA, Matin R, McCradden M, Oakden-Rayner L, Ordish J, Pearson R, Pfohl SR, Rostamzadeh N, Sapey E, Sebire N, Sounderajah V, Summers C, Treanor D, Denniston AK, Liu X. Tackling bias in AI health datasets through the STANDING Together initiative. Nat Med 2022;28:2232-2233. https://doi.org/10.1038/s41591-022-01987-w
63. Kadakia KT, Beckman AL, Ross JS, Krumholz HM. Leveraging Open Science to accelerate research. N Engl J Med 2021;384:e61. https://doi.org/10.1056/NEJMp2034518
64. Staniszewska S, Brett J, Simera I, Seers K, Mockford C, Goodlad S, Altman DG, Moher D, Barber R, Denegri S, Entwistle A, Littlejohns P, Morris C, Suleman R, Thomas V, Tysall C. GRIPP2 reporting checklists: tools to improve reporting of patient and public involvement in research. BMJ 2017;358:j3453. https://doi.org/10.1136/bmj.j3453
65. Camaradou JC, Hogg HD. Commentary: Patient perspectives on artificial intelligence; what have we learned and how should we move forward? Adv Ther 2023;40:2563-2572. https://doi.org/10.1007/s12325-023-02511-3
66. Finlayson SG, Beam AL, van Smeden M. Machine learning and statistics in clinical research articles-moving past the false dichotomy. JAMA Pediatr 2023;177:448-450. https://doi.org/10.1001/jamapediatrics.2023.0034
67. Sounderajah V, Ashrafian H, Golub RM, Shetty S, De Fauw J, Hooft L, Moons K, Collins G, Moher D, Bossuyt PM, Darzi A, Karthikesalingam A, Denniston AK, Mateen BA, Ting D, Treanor D, King D, Greaves F, Godwin J, Pearson-Stuttard J, Harling L, McInnes M, Rifai N, Tomasev N, Normahani P, Whiting P, Aggarwal R, Vollmer S, Markar SR, Panch T, Liu X. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. BMJ Open 2021;11:e047709. https://doi.org/10.1136/bmjopen-2020-047709
68. Mongan J, Moy L, Kahn CE. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell 2020;2:e200029. https://doi.org/10.1148/ryai.2020200029
69. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, Denniston AK, Faes L, Geerts B, Ibrahim M, Liu X, Mateen BA, Mathur P, McCradden MD, Morgan L, Ordish J, Rogers C, Saria S, Ting DS, Watkinson P, Weber W, Wheatstone P, McCulloch P. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med 2022;28:924-933. https://doi.org/10.1038/s41591-022-01772-9
70. Hawksworth C, Elvidge J, Knies S, Zemplenyi A, Petyko Z, Siirtola P, Chandra G, Srivastava D, Denniston A, Chalkidou A, Delaye J. Protocol for the development of an artificial intelligence extension to the Consolidated Health Economic Evaluation Reporting Standards (CHEERS) 2022. medRxiv [Preprint] 2023 Jun 1. https://doi.org/10.1101/2023.05.31.23290788
71. Rivera SC, Liu X, Chan AW, Denniston AK, Calvert MJ. Guide-

lines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension. BMJ 2020;370: m3210. https://doi.org/10.1136/bmj.m3210

72. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. BMJ 2020;370:m3164. https://doi.org/10.1136/bmj.m3164
73. Cacciamani GE, Chu TN, Sanford DI, Abreu A, Duddalwar V, Oberai A, Kuo CJ, Liu X, Denniston AK, Vasey B, McCulloch P, Wolff RF, Mallett S, Mongan J, Kahn CE, Sounderajah V, Darzi A, Dahm P, Moons KG, Topol E, Collins GS, Moher D, Gill IS, Hung AJ. PRISMA AI reporting guidelines for systematic reviews and meta-analyses on AI in healthcare. Nat Med 2023; 29:14-15. https://doi.org/10.1038/s41591-022-02139-w
74. Collins GS, Dhiman P, Ma J, Schlussel MM, Archer L, Van Calster B, Harrell FE, Martin GP, Moons KG, van Smeden M, Sperrin M, Bullock GS, Riley RD. Evaluation of clinical prediction models (part 1): from development to external validation. BMJ 2024;384:e074819. https://doi.org/10.1136/bmj-2023-074819
75. Riley RD, Archer L, Snell KI, Ensor J, Dhiman P, Martin GP, Bonnett LJ, Collins GS. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. BMJ 2024;384:e074820. https://doi.org/10.1136/bmj-2023-074820
76. Van Calster B, Steyerberg EW, Wynants L, van Smeden M. There is no such thing as a validated prediction model. BMC Med 2023;21:70. https://doi.org/10.1186/s12916-023-02779-w
77. Moher D, Schulz KF, Simera I, Altman DG. Guidance for developers of health research reporting guidelines. PLoS Med 2010;7:e1000217. https://doi.org/10.1371/journal.pmed.1000217
78. Collins GS, Moons KG. Reporting of artificial intelligence prediction models. Lancet 2019;393:1577-1579. https://doi.org/10.1016/S0140-6736(19)30037-6
79. Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, Logullo P, Beam AL, Peng L, Van Calster B, van Smeden M, Riley RD, Moons KG. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open 2021; 11:e048008. https://doi.org/10.1136/bmjopen-2020-048008
80. Gattrell WT, Logullo P, van Zuuren EJ, Price A, Hughes EL, Blazey P, Winchester CC, Tovey D, Goldman K, Hungin AP, Harrison N. ACCORD (ACcurate COnsensus Reporting Document): a reporting guideline for consensus methods in biomedicine developed via a modified Delphi. PLoS Med 2024; 21:e1004326. https://doi.org/10.1371/journal.pmed.1004326
81. Olczak J, Pavlopoulos J, Prijs J, Ijpma FF, Doornberg JN, Lundstrom C, Hedlund J, Gordon M. Presenting artificial intelligence, deep learning, and machine learning studies to clinicians and healthcare stakeholders: an introductory reference with a guideline and a Clinical AI Research (CAIR) checklist proposal. Acta Orthop 2021;92:513-525. https://doi.org/10.1080/17453674.2021.1918389
82. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, Arnaout R, Kohane IS, Saria S, Topol E, Obermeyer Z, Yu B, Butte AJ. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med 2020;26:1320-1324. https://doi.org/10.1038/s41591-020-1041-y
83. Hernandez-Boussard T, Bozkurt S, Ioannidis JP, Shah NH. MINIMAR (MINimum Information for Medical AI Reporting): developing reporting standards for artificial intelligence in health care. J Am Med Inform Assoc 2020;27:2011-2015. https://doi.org/10.1093/jamia/ocaa088
84. Scott I, Carter S, Coiera E. Clinician checklist for assessing suitability of machine learning applications in healthcare. BMJ Health Care Inform 2021;28:e100251. https://doi.org/10.1136/bmjhci-2020-100251
85. Schwendicke F, Singh T, Lee JH, Gaudin R, Chaurasia A, Wiegand T, Uribe S, Krois J. Artificial intelligence in dental research: checklist for authors, reviewers, readers. J Dent 2021;107: 103610. https://doi.org/10.1016/j.jdent.2021.103610
86. Sendak MP, Gao M, Brajer N, Balu S. Presenting machine learning model information to clinical end users with model facts labels. NPJ Digit Med 2020;3:41. https://doi.org/10.1038/s41746-020-0253-3
87. Stevens LM, Mortazavi BJ, Deo RC, Curtis L, Kao DP. Recommendations for reporting machine learning analyses in clinical research. Circ Cardiovasc Qual Outcomes 2020;13:e006556. https://doi.org/10.1161/CIRCOUTCOMES.120.006556
88. Kwong JC, McLoughlin LC, Haider M, Goldenberg MG, Erdman L, Rickard M, Lorenzo AJ, Hung AJ, Farcas M, Goldenberg L, Nguan C, Braga LH, Mamdani M, Goldenberg A, Kulkarni GS. Standardized Reporting of Machine Learning Applications in Urology: The STREAM-URO Framework. Eur

Urol Focus 2021;7:672-682. https://doi.org/10.1016/j.euf.2021.07.004

89. de Hond AA, Leeuwenberg AM, Hooft L, Kant IM, Nijman SW, van Os HJ, Aardoom JJ, Debray TP, Schuit E, van Smeden M, Reitsma JB, Steyerberg EW, Chavannes NH, Moons KG. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. NPJ Digit Med 2022;5:2. https://doi.org/10.1038/s41746-021-00549-7

90. Wolff RF, Moons KG, Riley RD, Whiting PF, Westwood M, Collins GS, Reitsma JB, Kleijnen J, Mallett S. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019;170:51-58. https://doi.org/10.7326/M18-1376

91. Moons KG, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, Reitsma JB, Kleijnen J, Mallett S. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med 2019;170:W1-W33. https://doi.org/10.7326/M18-1377

92. Ibrahim H, Liu X, Zariffa N, Morris AD, Denniston AK. Health data poverty: an assailable barrier to equitable digital health care. Lancet Digit Health 2021;3:e260-e265. https://doi.org/10.1016/S2589-7500(20)30317-4

93. McCradden MD, Joshi S, Mazwi M, Anderson JA. Ethical limitations of algorithmic fairness solutions in health care machine learning. Lancet Digit Health 2020;2:e221-e223. https://doi.org/10.1016/S2589-7500(20)30065-0

94. Mccradden M, Odusi O, Joshi S, Akrout I, Ndlovu K, Glocker B, Maicas G, Liu X, Mazwi M, Garnett T, Oakden-Rayner L, Alfred M, Sihlahla I, Shafei O, Goldenberg A. What's fair is… fair?: presenting JustEFAB, an ethical framework for operationalizing medical ethics and social justice in the integration of clinical machine learning: JustEFAB. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency; 2023 Jun 12-15; Chicago, USA. Association for Computing Machinery; 2023. p. 1505-1519. https://doi.org/10.1145/3593013.3594096

95. Thibault RT, Amaral OB, Argolo F, Bandrowski AE, Davidson AR, Drude NI. Open Science 2.0: towards a truly collaborative research ecosystem. PLoS Biol 2023;21:e3002362. https://doi.org/10.1371/journal.pbio.3002362

96. Riley RD, Tierney JF Stewart LA. Individual participant data meta-analysis: a handbook for healthcare research. Wiley; 2021.

eISSN: 2234-2591
Open Access

Guidelines

Ewha Med J 2025;48(3):e49
https://doi.org/10.12771/emj.2025.00661

The Ewha Medical Journal

대형 언어 모델을 활용한 연구를 위한 TRIPOD-LLM 보고 지침

The TRIPOD-LLM reporting guideline for studies using large language models

Jack Gallifant[1,2,3], Majid Afshar[4,29], Saleem Ameen[1,5,6,29], Yindalon Aphinyanaphongs[7,29], Shan Chen[3,8,29], Giovanni Cacciamani[9,10,29], Dina Demner-Fushman[11,29], Dmitriy Dligach[12,29], Roxana Daneshjou[13,14,29], Chrystinne Fernandes[1,29], Lasse Hyldig Hansen[1,15,29], Adam Landman[16,29], Lisa Lehmann[16,29], Liam G. McCoy[17,29], Timothy Miller[18,29], Amy Moreno[19,29], Nikolaj Munch[1,15,29], David Restrepo[1,20,29], Guergana Savova[18,29], Renato Umeton[21,29], Judy Wawira Gichoya[22,29], Gary S. Collins[23,24], Karel G. M. Moons[25,26], Leo A. Celi[1,27,28], Danielle S. Bitterman[3,8*]

For further information on the authors' affiliations, see Additional information.

요약

대형 언어 모델(large language model, LLM)의 활용이 의료 분야에서 빠르게 확대되면서, 표준화된 보고 지침의 필요성이 커지고 있다. 이 논문에서는 LLM을 활용한 연구를 위한 다변수 예측모델의 투명한 보고(TRIPOD-LLM) 지침을 제시하였다. TRIPOD-LLM은 기존 TRIPOD와 인공지능(artificial intelligence) 확장 지침을 기반으로 하며, 바이오 메디컬 분야에서 LLM이 가지는 고유한 도전 과제들을 반영하고 있다. 이 지침은 제목부터 논의까지 주요 내용을 포괄하는 19개 주요 항목과 50개 세부 항목으로 구성되어 있다. 다양한 LLM 연구설계와 작업에 적용할 수 있도록 모듈형 형식을 도입하였고, 모든 연구에 공통적으로 적용할 수 있는 14개 주요 항목과 32개 세부 항목을 포함한다. 이 지침은 신속한 델파이(Delphi) 과정과 전문가 합의를 거쳐 개발하였으며, 투명성과 인간 감독, 과업 특이적 성과(task-specific performance) 보고의 중요성을 강조한다. 또한 지침의 손쉬운 작성과 제출용 PDF 생성을 지원하는 인터랙티브 웹사이트(https://tripod-llm.vercel.app/)를 소개한다. TRIPOD-LLM은 '생명력 있는 문서'로서, 연구현장의 변화에 맞추어 지속적으로 개정될 예정이다. 이 지침을 통해 LLM 연구의 보고 수준을 높이고, 재현성과 임상 적용 가능성을 강화하는 데 기여하려고 한다.

서론

의료 분야에서 대형 언어 모델(large language model, LLM)의 도입은 계속 확대되고 있으며, 현재는 물론 미래에도 행정 및 의료

*Corresponding email: dbitterman@bwh.harvard.edu
Received: July 17, 2025 Revised: July 30, 2025 Accepted: July 30, 2025

[29] These authors contributed equally: Majid Afshar, Saleem Ameen, Yindalon Aphinyanaphongs, Shan Chen, Giovanni Cacciamani, Dina Demner-Fushman, Dmitriy Dligach, Roxana Daneshjou, Chrystinne Fernandes, Lasse Hyldig Hansen, Adam Landman, Lisa Lehmann, Liam G. McCoy, Timothy Miller, Amy Moreno, Nikolaj Munch, David Restrepo, Guergana Savova, Renato Umeton, and Judy Wawira Gichoya.

It is a Korean translation of the Gallifant J et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nature Med 2025;31:60-69 https://doi.org/10.1038/s41591-024-03425-5. The translation was done with the permission of the TRIPOD Group. Korean medical terminology is based on the English-Korean Medical Terminology 6th edition, available at: https://term.kma.org/index.asp.Korean translation was done by Sun Huh (https://orcid.org/0000-0002-8559-8640), Hallym University. The Korean proofreading was conducted by Yoon Joo Seo (https://orcid.org/0000-0002-0202-8352), InfoLumi, and the back-translation was performed by Jeong-Ju Yoo (https://orcid.org/0000-0002-7802-0381), Soonchunhyang University Bucheon Hospital. The back-translation was confirmed by Gary S. Collins (https://orcid.org/0000-0002-2772-2316), the one of the authors of the original TRIPOD-LLM statement.

서비스 제공 등 다양한 영역에서의 활용 가능성이 논의되고 있다. 대표적인 예로, 환자 커뮤니케이션 초안 생성, 의료문서 요약, 질의응답, 정보검색, 의료진단, 치료 권고, 환자교육, 의학교육 등이 있다[1-5]. LLM의 빠른 발전이 기존 규제 및 거버넌스 체계를 한계까지 밀어붙여, 이러한 복잡한 범용모델을 완전하게 반영하지 못하는 임시방편적인 해결책들이 나타나고 있다[6-8]. 더 나아가 LLM 개발이 가속화되면서 학술지 및 전문가 심사(peer review), 출판일정, 시의적절한 지침을 제공해야 하는 규제기관들도 어려움을 겪고 있다. 이에 대응하기 위해 연구자들은 프리프린트(preprint)를 빠르게 발표하고 있고, 보고할 때도 임시방편적 접근을 택하는 경우가 많다.

보고 지침(reporting guideline)은 연구의 표준화와 투명한 보고, 동료평가 절차를 위한 확장 가능한 틀을 제공한다. 중요한 예시로, 진단 및 예후예측모델 연구를 위한 최소 보고 기준을 확립하고자 2015년에 처음 도입된 transparent reporting of a multivariable model for individual prognosis or diagnosis (TRIPOD) 이니셔티브가 있다(https://www.tripod-statement.org) [9]. TRIPOD는 건강연구 보고의 질과 투명성 향상을 목표로 하는 국제적 노력인 enhancing the quality and transparency of health research (EQUATOR) 네트워크의 핵심 지침 중 하나이다[10]. TRIPOD는 여러 학술지가 폭넓게 지지 및 권고하고 있으며, 저자 안내문에도 자주 포함되어 있다. 이후 TRIPOD는 인공지능(artificial intelligence, AI)의 실질적 발전에 대응하여, 머신러닝(machine learning) 분야의 최신 모범 사례를 반영한 TRIPOD+AI로 업데이트되었다[11]. 또한 모델 수명 주기 전반에 걸친 AI 개발을 위한 보완적 지침들도 제시되어 왔다[12-14].

LLM은 AI 내에서도 독특한 특성을 지닌 새로운 영역으로, 기존 TRIPOD 지침이나 그 최신 확장판만으로는 완전히 다루기 어려운 고유한 도전 과제와 고려사항을 제시한다. 분류(classifier) AI 모델에서 생성형(generative) AI로의 전환이 이뤄지면서 이러한 특징이 더욱 부각되고 있다. 이에 이 논문에서는 이러한 충족되지 않은 요구를 해결하고, 빠르게 변화하는 연구현장에 유연하게 대응할 수 있도록 설계된 'TRIPOD-LLM statement'를 보고한다. 이 확장 지침은 원래 예측모델에 초점을 맞추었던 TRIPOD의 범위를 넘어, 진단부터 문서 요약까지 의료연구와 실무의 다양한 영역에 LLM이 미치는 광범위한 영향을 반영한다.

TRIPOD-LLM의 필요성

LLM은 자기회귀적(autoregressive) 구조를 가진 생성형 AI이다. 단순하게 말하면 앞선 단어들을 바탕으로 다음 단어를 예측하도록 학습된다는 의미이다. 그러나 이러한 기초적 학습만으로도 하나의 모델이 다양한 범위의 의료 관련 자연어 처리(natural language processing) 작업을 수행할 수 있다는 점이 확인되고 있다. 이러한 적응력은 주로 감독 학습 기반의 미세조정(supervised fine-tuning)이나 소수 예시 학습(few-shot learning) 방법을 통해 달성된다[15,16]. 이를 통해 LLM은 적은 예시만으로도 새로운 작업을 처리할 수 있다. 챗봇(chatbot) 솔루션(예: ChatGPT)은 LLM을 기반으로 하면서 두 가지 구성요소가 추가된다. 하나는 질의응답(질문에 대한 응답을 생성하는 instruction tuning 또는 supervised fine-tuning)이고, 다른 하나는 '얼라인먼트(alignment)'라고도 하는 선호도 랭킹이다. LLM과 챗봇에 특유한 이러한 방법론적 과정(예: 감독 미세조정에 사용된 하이퍼파라미터[hyperparameter]의 선택, 프롬프트[prompt] 설계의 복잡성, 모델 예측의 변동성, 자연어 출력 평가방법, 선호 기반 학습전략 등)은 현재의 보고 지침에서는 충분히 다루어지지 않고 있다. 이러한 과정은 모델의 신뢰성에 큰 영향을 미치므로, 별도의 구체적인 안내가 필요하다. 또한 LLM의 범용성과 생성형 특성은 기존 지침보다 더 세부적인 보고 지침을 요구한다. LLM은 특정 작업을 위해 훈련된 것이 아니며 훈련 데이터에 해당 작업이 반드시 포함되어 있지 않을 수 있으므로(기존 작업별 모델이 훈련 데이터에 해당 질병 유병률을 명시적으로 반영하는 경우와 달리), 신뢰성 있는 보고와 후속 안전성 확보를 위해서는 작업별로 특화된 지침이 필요하다.

생성형 출력물을 평가할 때 어떤 자동화된 혹은 인간 기반의 지표를 선택할 것인지에 대한 문제는 여전히 명확히 해결되지 않은 상태로, 현재 다양한 방법론이 적용되어 성능의 여러 측면을 평가하고 있다. 결과물이 완전히 비정형(unstructured) 텍스트이고 구조화된 라벨로 단순 환원이 불가능한 작업(예: 편지 생성, 요약 등)이라면 평가가 특히 복잡하다. 이런 경우 대부분의 자동 평가지표는 입력과 출력 텍스트 간의 중첩과 유사도에 초점을 맞추는데, 이는 생성된 텍스트의 사실적 정확성이나 적합성, 허위(hallucination)나 누락(omission)을 포착하지 못할 수 있다[17-19]. 이들 점수는 참조 텍스트와의 구조적·어휘적 유사성을 반영하지만, 이는 성능과 안전성을 종합적으로 평가하는 기준의 일부분에 불과하다. 인간의 텍스트 평가 또한 언어의 모호성, 임상 과제의 불확실성 등으로 인해 주관적일 수밖에 없다. 특히 의학 분야에서는 정답이 하나로 규정되지 않는 경우가 많고, 무작위적(aleatoric)이고 지식적(epistemic)인 불확실성이 모두 존재한다. 따라서 성능이 어떻게 평가되었는지를 보고하기 위한 구체적인 세부 지침이 필요하다. 본 논문에서는 LLM과 챗봇을 모두 포괄하여 'LLM'으로 지칭한다. Table 1은 의료 분야에 적용 가능한 주요 작업유형을 정리하고, 관련 선행연구의 정의 및 예시를 제시한다[5,6,20-30].

LLM이 도입됨에 따라 환각(hallucination), 누락(omission), 신뢰성, 설명 가능성, 재현성, 프라이버시, 편향의 하위 전파 등과 같은 새로운 복잡성이 나타나고 있는데, 이는 임상 의사결정과 환자 진료에 부정적 영향을 미칠 수 있다[20,21,31-35]. 더불어 전자의무기록(electronic health record) 업체, 기술기업, 의료기관 간의 협력이 확대되면서, 실제 적용 일정은 현행 규제의 대응속도를 훨

Table 1. TRIPOD-LLM 가이드라인의 모듈형 연구설계 및 LLM 과업 범주

과업(task)	정의(definition)	예시(example)
연구 설계(research design)		
De novo LLM 개발	새로운 언어 모델을 처음부터 구축하거나, 기존 기본(base) 모델을 상당 부분 미세 조정하여 새로운 기능을 개발하거나 새로운 작업에 적응시키는 작업	병원 임상 데이터로 새로운 LLM을 사전 학습(pretraining)하는 연구[22]
LLM 방법(LLM methods)	언어 모델의 새로운 아키텍처, LLM을 이해하기 위한 새로운 계산방법, LLM 평가를 위한 새로운 방법, LLM 프롬프트 최적화를 위한 새로운 방법 등에 중점을 두는 정량적 또는 이론적 연구	의료 분야에서 retrieval-augmented generation LLM 프레임워크를 연구하는 연구[23]
LLM 평가(LLM evaluation)	기존 LLM의 효율성, 정확성, 특정 의료작업에의 적합성을 평가하거나, 사용 시 발생하는 위험과 편향을 평가하는 작업	기존 LLM에서 편향된 진단추론을 조사하는 연구[20]
의료환경에서의 LLM 평가(LLM evaluation in healthcare settings)	임상 워크플로우 내에서 LLM이 통합되어 실제로 사용될 때, 임상적 · 행정적 · 인력 관련 결과(outcomes) 측면의 영향과 통합에 초점을 두고 평가하는 작업	병원 입원환자에서 실시간(real-time) 예후예측을 위해 LLM을 배치하여 그 성능을 보고하는 연구[6]
LLM 과업(LLM task)		
텍스트 처리(text processing)	토큰화(tokenization), 구문 분석(parsing), 엔터티 인식(entity recognition) 등, 텍스트 데이터를 조작하거나 하위 수준으로 처리하는 작업을 포함하되, 이에 국한되지는 않음	LLM 기반 명명 엔터티 인식(named entity recognition) 방법을 연구하는 연구[24]
분류(classification)	텍스트 데이터에 미리 정의된 라벨을 할당하는 작업	임상노트에서 1개 이상의 사회적 결정요인(social determinants of health)이 언급되었는지 여부를 판단하기 위해 LLM을 미세 조정하는 연구[25]
장문 질의응답(long-form question answering)	복잡한 질의에 대해 여러 문서 또는 근거를 바탕으로 상세한 답변을 제공하는 작업(객관식 질의응답[multichoice Q&A]은 '분류[classification]'에 포함됨)	기존 LLM이 환자 포털 메시지에 답변하는 능력을 조사하는 연구[21]
정보 검색(information retrieval)	대규모 데이터 세트에서 특정 질의에 따라 관련 정보를 추출하는 작업. 문헌 리뷰, 환자 병력 조회 등에 적용됨	트랜스포머 모델을 훈련하여 사용자의 질의에 적합한 생의학 논문을 검색하도록 한 연구[26]
대화형 에이전트(conversational agent, 챗봇)	사용자와 대화를 이어가는 작업. 환자 상호작용, 건강상담, 의료진의 가상 비서 등으로 활용됨	LLM 기반 챗봇 접근 가능 여부가 임상의의 진단추론에 미치는 영향을 조사하는 연구[27]
문서 생성(documentation generation)	임상 데이터, 녹취, 기록 등을 바탕으로 자동으로 의료문서를 생성하는 작업	임상환경 녹음에서 자동 생성된 임상노트의 품질을 평가하는 연구[5]
요약 및 단순화(summarization and simplification)	방대한 텍스트 문서를 요약하거나, 쉽게 이해할 수 있도록 내용을 단순화하는 작업. 환자 교육, 의무기록 요약 등에서 유용함	LLM이 퇴원 요약(discharge summary)을 환자 친화적 평이문으로 변환하는 능력을 평가하는 연구[28]
기계 번역(machine translation)	한 언어의 텍스트를 다른 언어로 변환하는 작업	번역에 특화된 소형 언어모델과 범용 LLM이 스페인어-영어 생의학 텍스트를 번역하는 능력을 비교한 연구[29]
결과 예측(outcome forecasting)	과거 데이터를 기반으로 미래의 의료결과를 예측하는 작업. 예후 평가나 치료효과 연구에 활용됨	LLM이 중환자실 입원 환자의 병원 외 사망률을 예측하는 능력을 연구한 논문[30]

TRIPOD, transparent reporting of a multivariable model for individual prognosis or diagnosis; LLM, large language model.

씬 앞지르고 있다[8,36]. LLM의 안전한 사용과 투명성 제고를 위해서는 개발 및 보고의 표준화가 필수적이다. 이는 일관성, 신뢰성, 검증 가능성 확보를 위한 것으로, 타 과학 분야에서 확립된 임상평가와 유사한 수준의 기준이 필요하다[37-39].

방법

TRIPOD-LLM 지침은 LLM을 개발, 튜닝 또는 평가하는 모든 의료 응용 또는 맥락에서 연구 보고를 안내하기 위해 작성되었으며, 기존 TRIPOD 지침 개발과정과 동일한 절차를 따랐다. 본 분야에 시의적절한 보고 지침이 필요하다는 점을 고려하여 신속한 델

Box 1. 용어 해설(glossary of terms)

아래 정의와 설명은 TRIPOD-LLM의 특정 맥락 및 본 가이드라인에서의 용례에 한정된 것으로, 다른 연구 분야에는 그대로 적용되지 않을 수 있다.

Attention mechanism(어텐션 메커니즘): 신경망에서 출력의 각 부분을 생성할 때 입력의 서로 다른 부분에 주의를 집중할 수 있도록 하여, 시퀀스 데이터 내 장거리 의존성 처리를 가능하게 하는 핵심 요소.

Chain-of-thought prompting(사고 흐름 프롬프트): 모델이 복잡한 추론 과제를 단계별 사고과정으로 분해하여 처리하도록 유도하는 프롬프트 기법으로, 논리적·수리적 문제 해결력 향상에 도움을 준다.

Confabulation(혼동 생성): Hallucination(환각)의 대체용어로, 의도하지 않게 허위정보를 생성하는 현상을 의미한다.

Data leakage(데이터 누출): 모델학습 또는 미세조정 과정에서 테스트 데이터를 사용하는 것으로, 실제 성능보다 과대평가되는 결과를 초래한다.

Decoder(디코더): 벡터화된 입력 데이터를 다시 텍스트 시퀀스로 변환하는 모델 구성요소.

Autoregressive model(자기회귀모델): 시퀀스 내 앞선 요소를 바탕으로 다음 요소(예: 문장의 다음 단어)를 예측하는 트랜스포머 기반 모델. 최신 LLM (생성형 사전학습 변환기 등)은 대부분 자기회귀모델이다.

Embedding(임베딩): 텍스트를 고차원 벡터 공간에 표현하여, 의미적으로 유사한 단어가 비슷한 벡터로 나타나게 하는 방식('벡터' 참조).

Encoder(인코더): 입력 데이터를 벡터화하거나 모델이 이해할 수 있는 표현으로 변환하는 모델의 구성요소.

Encoder-decoder(인코더-디코더): 인코더와 디코더를 결합하여 입력 데이터를 출력으로 변환하는 모델 아키텍처 프레임워크.

Few-shot learning(소수 예시 학습): 매우 적은 수의 예시만으로 모델이 작업을 효과적으로 수행할 수 있도록 하는 학습방법. 예시 수에 따라 one-shot learning 등으로 명명되기도 한다.

Fine-tuning(미세조정): 사전 학습된 모델을 소규모 도메인 특화 데이터 세트로 추가 학습시켜 특정 작업에 특화하는 과정.

GPT: 자연어 이해 및 생성을 위한 자기회귀 트랜스포머 기반 모델 계열. 문장 내 다음 단어 예측을 위한 사전 학습을 수행함.

Hallucination(환각): 언어모델이 입력 데이터와 무관하거나 관련성이 적은 텍스트를 생성하는 현상. 허위, 부정확한 내용이 포함될 수 있다.

In-context learning(문맥 내 학습): 추론 단계에서 프롬프트에 예시를 제공함으로써 모델이 새로운 작업을 학습하는 능력.

Instruction tuning(지시 조정): 자연어 지시문과 바람직한 출력 쌍의 데이터 세트를 활용하여 모델이 다양한 지시문을 잘 따르도록 미세 조정하는 방법.

Prompt(프롬프트): LLM에게 응답을 유도하기 위해 입력하는 질의 또는 지시문.

Reinforcement learning(강화학습): LLM 개발에서 흔히 사용되는 머신러닝 기법으로, 행동에 대해 보상을 주어 모델이 인간 선호에 맞는 출력을 내도록 학습한다.

Prompt engineering(프롬프트 엔지니어링): 원하는 출력을 얻기 위해 프롬프트를 설계·최적화하는 과정. 반복적 프롬프트, 예시 포함, 사고 흐름 프롬프트 등이 포함된다.

Retrieval-augmented generation(검색 기반 생성): 외부 지식 베이스에서 정보를 검색해 생성과정에 결합, LLM이 최신 또는 도메인 특화 정보를 활용해 텍스트를 생성하도록 하는 방법.

Temperature(온도 파라미터): Softmax 적용 전 로짓을 조정하여 예측의 무작위성을 제어하는 파라미터. 값이 높을수록 생성 텍스트의 다양성이 커진다.

Tokenization(토크나이즈): 텍스트를 단어, 어절, 구 등 작은 단위(token)로 분할해 자연어 처리작업을 용이하게 하는 과정.

Transformer(트랜스포머): NLP 분야의 혁신적 신경망 아키텍처. 자기 어텐션 메커니즘(self-attention)을 통해 시퀀스를 병렬 처리하고, 복잡한 종속관계도 효과적으로 포착한다.

Vector(벡터): 데이터의 수치적 표현. LLM에서는 각 토큰(단어 조각)의 벡터가 주변 단어에 의해 영향을 받는 '문맥 임베딩(contextual embedding)'으로 나타난다.

Zero-shot learning(제로샷 학습): 모델의 이해 및 일반화 능력에 기반해 명시적으로 학습한 적 없는 작업도 올바르게 수행하는 능력.

파이(Delphi) 과정을 적용하였고, 이를 생명력 있는(living) 지침 방식과 결합하였다. 관련 용어에 대한 정의는 부록(Box 1)으로 수록하였다.

가이드라인 개발을 위한 운영위원회를 구성하고, natural language processing (NLP), AI, 의료정보학 등 다양한 전문성과 경험을 갖춘 전문가 패널과 함께 작업하였다. (본 논문 저자의 두 그룹 내 역할은 Author contributions에 기술했다.) 이 지침은 2024년 5월 2일, EQUATOR Network에 개발 중인 보고 지침으로 등록되었다(https://www.equator-network.org).

윤리 선언

본 연구는 2024년 3월 26일 MIT 인간대상 실험 윤리위원회(COUHES IRB)로부터 면제 승인을 받았다(exempt ID: E-5705). 델파이 설문 참여자는 설문 응답 전 전자 동의서를 제출하였다.

후보 항목 목록 도출

TRIPOD-2015, TRIPOD+AI 가이드라인(https://www.tripod-statement.org) 및 LLM 보고 지침 관련 문헌을 참고하여 초기 후보 항목 목록을 작성하였다[9,11,37,40]. 운영위원회와 전문가 패널은 추가 문헌조사를 통해 이 목록을 확장하였고, 최종적으

로 제목, 초록, 서론, 방법, 결과, 논의, 기타 항목 등 총 64개의 고유 항목으로 표준화하였다.

패널 모집

델파이 참여자는 운영위원회가 관련 논문 저자와 개인 추천, 그리고 델파이 참여자가 추천한 전문가를 포함하여 선별하였다. 운영위원회는 지리적 및 학문적 다양성을 반영하여 연구자(통계학자, 데이터 과학자, 역학자, 머신러닝 전문가, 임상의, 윤리학자), 의료전문가, 학술지 편집자, 연구비 지원자, 정책 입안자, 의료규제자, 환자 옹호단체 등 주요 이해관계자를 포함시켰다. 참여자 수에는 최소 표본크기 제한을 두지 않았다. 운영위원회 구성원이 각 참여자의 전문성 또는 경험을 확인하였다. 이후 이메일을 통해 설문 참여를 요청하였으며, 참여자에게 금전적 보상이나 선물은 제공하지 않았다.

델파이 과정

설문은 개별 응답이 가능하도록 영어로 설계되었으며, Google Forms (Google LLC)를 통해 온라인으로 배포되었다. 모든 응답은 익명으로 처리하였고, 이메일 또는 식별 정보는 수집하지 않았다. 참여자에게 각 항목을 '생략 가능,' '포함 가능,' '포함 권장,' '포함 필수'로 평가하도록 요청하였다. 이는 이전 TRIPOD 가이드라인에도 적용한 방식이다[9]. 모든 항목에 대해 의견을 남기거나 새로운 항목을 제안할 수 있었다. 자유 응답은 D.S.B.와 J.G.가 수합 및 분석하였으며, 그 결과를 토대로 항목의 문구 수정, 통합, 신규 항목 제안을 진행하였다. 모든 운영위원회 구성원도 델파이 설문에 참여할 수 있도록 초대되었다.

1차 참여자

1차 설문은 2024년 3월 1일부터 4월 23일까지 실시하였고, 설문 링크는 56명에게 발송하였다. 이 중 26명이 설문을 완료하였다. 설문 참여자는 9개국 출신으로, 북미 14명, 유럽 5명, 아시아 2명, 남미 1명, 오스트랄라시아 1명이었다. 3명은 정보를 제공하지 않았다. 참여자는 주된 업무 분야를 복수로 선택할 수 있었는데, 26명 중 20명(77%)이 AI, 머신러닝, 임상정보학, NLP 분야를, 14명이 의료 분야를 주요 분야로 선택하였다.

합의 회의

4월 22일과 24일, D.S.B.와 J.G.의 주재하에 모든 운영위원회 및 전문가 패널 구성원을 초대하여 온라인(Zoom; Zoom Video Communications Inc.)으로 합의 회의를 진행하였다. 회의 녹화본과 회의록은 불참자를 위해 즉시 배포하여 회의 이후에도 의견 제출이 가능하도록 하였다. 각 질문에 대한 응답과 자유 의견을 차례로 검토하였고, '포함 필수'에 대해 50% 미만의 지지를 받은 항목은 따로 표시하여 포함의 중요성에 대해 심도 있게 논의하였다. 모든 항목에서 합의(consensus)가 이뤄졌으며, 제3자의 개입은 필요하지 않았다. 합의에 도달할 때까지, 패널 중 추가 의견이나 이견이 없을 때까지 항목을 논의하였다. 논의 녹취록은 개인식별 정보와 민감 정보를 제거한 후 Supplementary Information에 공개하여 투명성을 확보하였다.

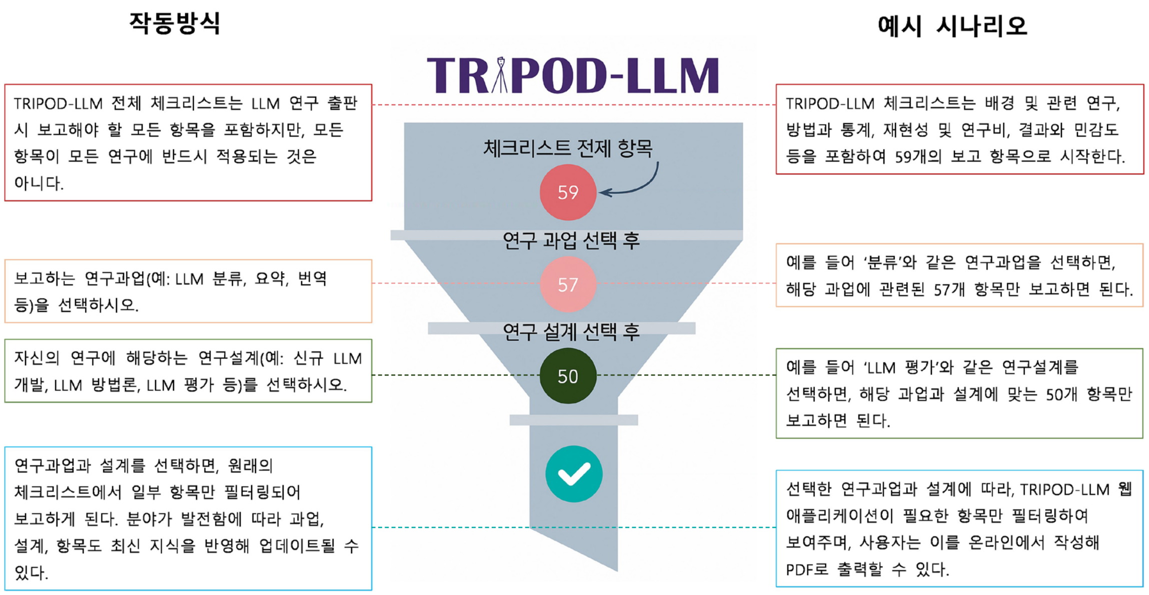
LLM 활용 분야의 방대함을 고려하여, '연구설계(Research Design)'와 'LLM 과업(LLM Task)' 아래 추가 하위 범주로 항목을 그룹화하는 모듈형 방식을 도입하였다. 이 방안은 합의 회의에서 채택되었고, 최종 분류는 운영위원회의 승인을 받았다.

LLM의 개발, 튜닝, 평가, 적용 등 다양한 단계와 여러 의료과업에서의 활용을 적절히 반영하기 위해, 항목은 (1) 연구설계와 (2) LLM 과업을 기준으로 분류하였다(Fig. 1). 연구설계 범주는 *de novo* LLM 개발, LLM 방법(미세조정, 프롬프트 엔지니어링, 아키텍처 수정 등), LLM 내재 평가, 헬스케어 환경에서의 LLM 평가 등으로 구분된다. LLM 과업 범주는 저수준 텍스트 처리(품사 태깅, 관계 추출, 명명 엔티티 인식 등), 분류(진단 등), 장문의 질의응답, 대화형 에이전트, 문서 생성, 요약/단순화, 기계번역, 결과 예측(예: 예후 등)이다. 각 항목은 복수의 설계 및 과업 범주에 해당할 수 있고, 한 연구에 둘 이상의 설계 및 과업이 포함될 수도 있다. 연구에 포함된 모든 설계 및 과업에 해당하는 항목은 반드시 보고해야 한다. 각 설계 및 과업 범주에 대한 정의와 예시는 Table 1에 제시하였다. 이러한 분류가 완벽하지 않으며, 설계 및 과업 간 중복이 존재할 수 있음을 인정한다.

TRIPOD-LLM 지침

TRIPOD-LLM은 LLM을 개발, 튜닝, 프롬프트 엔지니어링하거나 평가하는 연구에서 필수적인 항목들을 적절하게 보고할 수 있도록 구성된 체크리스트를 포함한다(Table 2).

Box 2에는 TRIPOD-2015 및 TRIPOD+AI에 추가되거나 변경된 주요 내용이 요약되어 있으며, Box 1에는 주요 정의가 제시되어 있다. TRIPOD-LLM 체크리스트는 제목(1개 항목), 초록(1개 항목), 서론(2개 항목), 방법(8개 항목), 오픈 사이언스 실천(1개 항목), 환자 및 대중의 참여(1개 항목), 결과(3개 항목), 논의(2개 항목) 등 총 19개 주요 항목으로 구성되어 있다. 이 주요 항목들은 50개의 세부 항목으로 세분화된다. 이 중 14개 주요 항목과 32개 세부 항목은 모든 연구설계와 LLM 작업에 공통으로 적용되고, 나머지 5개 주요 항목과 18개 세부 항목은 특정 연구설계나 LLM 작업 유형에만 적용된다. 방법(Methods)에서 논의된 바와 같이, TRIPOD-LLM 지침은 다양한 LLM 연구유형에 대응할 수 있도록 모



TRIPOD-LLM 전체 체크리스트는 LLM 연구 출판 시 보고해야 할 모든 항목을 포함하지만, 모든 항목이 모든 연구에 반드시 적용되는 것은 아니다.

보고하는 연구과업(예: LLM 분류, 요약, 번역 등)을 선택하시오.

자신의 연구에 해당하는 연구설계(예: 신규 LLM 개발, LLM 방법론, LLM 평가 등)를 선택하시오.

연구과업과 설계를 선택하면, 원래의 체크리스트에서 일부 항목만 필터링되어 보고하게 된다. 분야가 발전함에 따라 과업, 설계, 항목도 최신 지식을 반영해 업데이트될 수 있다.

TRIPOD-LLM 체크리스트는 배경 및 관련 연구, 방법과 통계, 재현성 및 연구비, 결과와 민감도 등을 포함하여 59개의 보고 항목으로 시작한다.

예를 들어 '분류'와 같은 연구과업을 선택하면, 해당 과업에 관련된 57개 항목만 보고하면 된다.

예를 들어 'LLM 평가'와 같은 연구설계를 선택하면, 해당 과업과 설계에 맞는 50개 항목만 보고하면 된다.

선택한 연구과업과 설계에 따라, TRIPOD-LLM 웹 애플리케이션이 필요한 항목만 필터링하여 보여주며, 사용자는 이를 온라인에서 작성해 PDF로 출력할 수 있다.

Fig. 1. TRIPOD-LLM 워크플로우. TRIPOD-LLM 체크리스트 워크플로우는 총 59개의 보고 항목으로 시작하며, 연구과업(예: 분류, 요약)과 연구설계(예: LLM 평가) 선택에 따라 필요한 항목 수가 점차 줄어든다. 두 가지 모두 선택한 후에는, 보고에 필요한 항목만 필터링된 목록이 생성된다. TRIPOD, transparent reporting of a multivariable model for individual prognosis or diagnosis; LLM, large language model.

듈형 형식을 도입하였다(Table 1). 일부 항목은 특정 연구설계나 LLM 작업유형에만 해당된다. 이러한 설계 및 작업 범주는 폭넓지만 상호 배타적인 것은 아니며, 연구의 맥락에 따라 달라질 수 있고 LLM의 적용이 진화함에 따라 변화가 필요할 수 있다. 또한 LLM 기반 연구의 학술지 또는 학회 초록을 위한 별도의 체크리스트를 포함하고 있으며, 기존 TRIPOD+AI for abstracts 지침도 개정되어(TRIPOD-LLM for abstracts) (Table 3), 새로운 내용을 반영하고 TRIPOD-LLM과의 일관성을 유지하였다[18].

TRIPOD-LLM에 포함된 권고사항은 LLM 기반 연구가 어떻게 수행되었는지 완전하고 투명하게 보고하기 위한 것으로, LLM을 개발하거나 평가하는 방법 자체를 규정하지는 않는다. 이 체크리스트는 연구의 질을 평가하는 도구가 아니며, CANGARU (ChatGPT, generative artificial intelligence and natural large language models for accountable reporting and use) [41] 및 CHART (Chatbot Assessment Reporting Tool) [40] 역시 생성형 AI와 챗봇에 초점을 맞춘 보완적 지침이다.

TRIPOD 공식 웹사이트(https://www.tripod-statement.org) 외에도, 연구설계와 작업에 따라 필요한 질문을 제공하는 인터랙티브 웹사이트(https://tripod-llm.vercel.app/)가 함께 개발되어 작성의 용이성을 높였다. 이 사이트에서는 제출에 적합한 최종 PDF를 생성할 수 있다. TRIPOD-LLM 체크리스트의 (입력 가능한) 작성용 템플릿은 https://www.tripod-statement.org에서 다운로드할 수 있다. TRIPOD-LLM과 후속 지침 발표와 관련된 소식, 공지, 정보 등은 TRIPOD-LLM 웹사이트와 TRIPOD 공식 웹사이트 (https://www.tripod-statement.org)에서 확인할 수 있다.

임상 및 병원 운영작업을 위해 LLM의 사전학습(pretraining), 미세조정(fine-tuning), 후향 평가(retrospective evaluation), 임상 적용(clinical deployment)에 관한 이전에 발표된 연구를 대상으로 작성한 TRIPOD-LLM 체크리스트의 완성 예시는 Supplement 1에 제시되어 있다[6]. 또한 이해를 돕기 위해 작성용 체크리스트는 Supplement 2으로, 새로운 항목에 대한 설명과 보충 문서는 Supplement 3으로 제공하였다.

생명력 있는(living) 문서로서의 TRIPOD-LLM 지침

이 분야의 급속한 발전속도와 의료종사자 및 환자와의 상호작용 시점을 고려하여, (생명)과학 및 기타 의료 분야에서 LLM을 적시에 활용할 수 있도록 TRIPOD-LLM 지침을 신속히 마련하기로 결정하였다. 이 지침은 사용자 테스트를 통한 개선, 분야 발전에 따른 업데이트, 새로운 표준의 도입 및 정기적인 검토를 용이하게 하기 위하여, 생명력 있는(living) 문서형태로 설계되어 인터랙티브 웹사이트에 게시되어 있다. 따라서 향후 보고 권고사항의 지속적인 변화가 예상되므로, 사용자들은 항상 https://tripod-llm.vercel.app/에 게시된 최신 버전의 가이드라인을 참고하는 것이 좋다.

TRIPOD-LLM 지침을 유연한 변화가 가능한(living) 방식으로 개발한 것은, 변화하는 근거에 기반하여 최신 권고를 제공하기 위해 고안된 living systematic reviews [42,43] 및 임상진료지침(clin-

Table 2. TRIPOD-LLM 점검표

섹션(section)	항목(item)	설명(description)	연구설계(research design)	LLM 과업(LLM task)
제목(Title)	1	해당 연구가 LLM의 개발, 미세조정 또는 성능 평가임을 밝히고, 수행작업, 대상 집단, 예측하고자 하는 결과를 명시할 것	All	All
초록(Abstract)	2	TRIPOD-LLM for abstracts 참조	All	All
서론(Introduction)				
배경(Background)	3a	의료 맥락/활용 사례(예: 행정, 진단, 치료, 임상 워크플로우)와 LLM 개발 또는 평가의 근거를 설명하고, 기존 접근법 및 모델을 인용할 것	All	All
	3b	대상 집단과 케어 경로 내에서 LLM의 의도된 사용, 현재 표준 실무의 의도된 사용자(예: 의료전문가, 환자, 대중, 행정가)를 포함하여 설명할 것	E, H	All
목적(Objectives)	4	연구목적을 명시하되, 연구가 LLM의 초기 개발, 미세조정, 검증 중 어떤 단계(혹은 여러 단계)에 대한 것인지 포함할 것	All	All
방법(Methods)				
데이터(Data)	5a	학습, 튜닝, 평가 데이터 세트의 출처를 개별적으로 기술하고, 해당 데이터를 사용한 근거를 제시할 것(예: 웹 코퍼스[corpus], 임상연구/시험 데이터, EHR 데이터 등)	All	All
	5b	관련 데이터 항목을 기술하고, 분포 등 데이터 세트의 정량적 · 정성적 특성과 기타 관련 설명자(예: 출처, 언어, 국가 등) 제공	All	All
	5c	개발(학습, 미세조정, reward modeling) 및 평가 데이터 세트에서 사용한 텍스트의 가장 오래된 날짜와 최신 날짜를 명확히 제시할 것	All	All
	5d	데이터 전처리 및 품질 검사방법을 기술하되, 이 과정이 텍스트 코퍼스, 기관, 사회인구학적 집단 간에 동일했는지 포함할 것	All	All
	5e	결측 및 불균형 데이터 처리방법과 데이터 제외 사유를 명확히 기술할 것	All	All
분석방법(Analytical methods)	6a	LLM 이름, 버전, 최종 학습날짜를 보고할 것	All	All
	6b	LLM 개발과정(아키텍처, 학습, 미세조정 절차, 얼라인먼트 전략[예: 강화학습, 직접 선호도 최적화]과 그 목표[예: 유용성, 정직성, 무해성 등])에 대한 세부사항 보고	M, D	All
	6c	LLM을 활용한 텍스트 생성과정, 프롬프트 엔지니어링(출력 일관성 포함), 추론 설정(예: 시드 값, 온도, 최대 토큰 길이, 페널티 등) 상세 보고	M, D, E	All
	6d	LLM의 초기 및 후처리 출력값 명시(예: 확률, 분류, 비정형 텍스트 등)	All	All
	6e	분류(classification) 관련 세부 내용 및 확률 산출방법과 임계값 식별방법 포함 시 상세 설명	All	C, OF
LLM 출력(LLM output)	7a	생성결과의 품질(일관성, 관련성 등) 평가지표 포함	All	QA, IR, DG, SS, MT
	7b	배포 시점의 하위 과업 결과지표의 적합성과, 해당 용도와 인간 평가와의 상관관계 보고	E, H	All
	7c	결과 정의, LLM 예측 산출방식(예: 공식, 코드, 객체, API), 폐쇄형 LLM의 추론날짜, 평가지표 명확화	E, H	All
	7d	결과 평가에 주관적 해석이 필요할 경우, 평가자의 자격, 제공된 지침, 평가자 인구통계 정보, 평가자 간 합의 포함	All	All
	7e	성능 비교 시 기준(LLM, 인간, 기타 벤치마크/표준) 명시	All	All
주석(Annotation)	8a	주석을 작성한 경우, 텍스트 라벨링 방식, 구체적 가이드라인 및 예시 포함해 기술	All	All
	8b	주석자 수, 각 데이터 세트별 다중 주석 비율, 주석자 간 합의 등 포함	All	All
	8c	주석자 배경 및 경험, 라벨링에 관여한 모델 특성 등 정보 제공	All	All

(Continued on the next page)

Table 2. Continued

섹션(section)	항목 (item)	설명(description)	연구설계 (research design)	LLM 과업 (LLM task)
프롬프트(Prompting)	9a	프롬프트 관련 연구일 경우, 프롬프트 설계, 선별, 선정과정 상세 기술	All	All
	9b	프롬프트 개발에 사용된 데이터 명시	All	All
요약(Summarization)	10	요약 전 데이터 전처리 방법을 기술	All	SS
지시 조정/얼라인먼트(Instruction tuning/alignment)	11	해당 전략 사용 시, 평가에 사용된 지침, 데이터, 인터페이스 및 평가 집단 특성 명시	M, D	All
연산 자원(Compute)	12	연구 수행에 필요한 연산 자원 또는 그 대체지표(예: 사용한 기계 수, 소요시간, 비용, 추론시간, 초당 부동소수점 연산 횟수[FLOPS] 등) 보고	M, D, E	All
윤리 승인(Ethical approval)	13	연구를 승인한 IRB 또는 윤리위원회 명칭, 참여자 동의 또는 동의 면제 여부 명시	All	All
오픈 사이언스(Open science)	14a	연구 자금 출처와 연구비 제공자의 역할 명시	All	All
	14b	모든 저자의 이해상충 및 재정 공개 선언	All	All
	14c	연구 프로토콜 공개 위치 명시 또는 미작성 명시	H	All
	14d	연구 등록정보(등록기관, 등록번호) 명시 또는 미등록 명시	H	All
	14e	연구 데이터 이용 가능성 상세 안내	All	All
	14f	연구결과 재현을 위한 코드 이용 가능성 상세 안내	All	All
공공 참여(Public involvement)	15	설계, 수행, 보고, 해석, 결과 확산 등 과정에서 환자 및 공공 참여 내역 또는 미참여 사실 명시	H	All
결과(Results)				
참여자(Participants)	16a	환자/EHR 데이터 사용 시, 텍스트/EHR/환자 데이터의 흐름, 문서/질문/참여자 수(결과 보유 · 미보유 구분), 추적기간 등 기술	EH	All
	16b	환자/EHR 데이터 사용 시, 전체 및 각 데이터 출처/설정/개발 · 평가 분할별 특성, 주요 날짜, 주요 특성, 표본크기 등 보고	EH	All
	16c	임상결과 포함 LLM 평가 시, 개발 및 평가 데이터 간 주요 임상 변수 분포 비교(가능 시)	EH	All
	16d	환자/EHR 데이터 사용 시, 각 분석(LLM 개발, 하이퍼파라미터 튜닝, 평가 등)별 참여자 및 결과 발생건수 명시	EH	All
성능(Performance)	17	사전 지정한 평가지표(7a) 및/또는 인간 평가(7d)에 따라 LLM 성능 보고	All	All
LLM 업데이트(LLM updating)	18	해당되는 경우, LLM 업데이트 결과와 이후 성능 보고	All	All
논의(Discussion)				
해석(Interpretation)	19a	주요 결과의 전반적 해석, 목표와 선행연구 맥락에서의 공정성 이슈 포함	All	All
한계(Limitations)	19b	연구 한계와 이에 따른 편향, 통계적 불확실성, 일반화 가능성에 미치는 영향 논의	All	All
맥락 내 LLM 활용성(Usability of the LLM in context)	19c	지정된 작업과 도메인 맥락에서 데이터를 사용하는 데 있어, 표현(representation), 결측(missingness)	E, H	All
	19d	평가대상 적용사례의 의도된 사용 목적, 입력, 최종 사용자, 자율성/인간 감독 수준 정의	E, H	All
	19e	해당 시, LLM 적용 시 저품질 또는 이용 불가 입력 데이터의 평가 · 처리방법, 실제 임상에서의 LLM 활용성 기술	E, H	All
	19f	해당 시, 사용자가 입력 데이터 처리나 LLM 사용에 관여해야 하는지와 필요한 전문성 수준 명시	E, H	All
	19g	향후 연구의 다음 단계, LLM의 적용 가능성 및 일반화 가능성 중심으로 논의	All	All

기존 LLM을 활용하는 연구의 경우 사용자는 원 개발자가 제공한 보고 가능한 정보에 대한 참고문헌을 반드시 포함해야 하며, 해당 정보가 제공되지 않은 경우에는 이를 명시해야 한다.

TRIPOD, transparent reporting of a multivariable model for individual prognosis or diagnosis; LLM, large language model; HER, electronic health record; API, application programming interface; FLOPS, floating-point operations per second; IRB, Institutional Review Board; E, LLM evaluation; H, LLM evaluation in healthcare settings; M, LLM methods; D, *de novo* LLM development; C, classification; OF, outcome forecasting; QA, long-form question answering; IR, information retrieval; DG, document generation; SS, summarization and simplification; MT, machine translation.

Box 2. TRIPOD-LLM에서 TRIPOD-2015 및 TRIPOD+AI와 달라진 주요 내용 및 추가 사항

LLM 보고를 위한 신규 체크리스트 도입: LLM의 고유한 특성과, 기존 AI 및 예측모델과 구별되는 특정 방법론을 반영하여 LLM 보고에 특화된 별도의 체크리스트가 개발되었다.

생명력 있는(living) 가이드라인: 이 체크리스트는 생명력 있는 문서로 설계되어, 문헌 검토와 커뮤니티의 의견을 반영하여 정기적으로 업데이트된다. 이 방식은 본 분야의 빠른 발전속도를 반영하기 위한 것으로, 신속한 버전 관리, 사용자 테스트를 통한 개선, 적시 업데이트가 가능하도록 하였다.

과업 특이적(task-specific) 지침 신설: 체크리스트에는 의료 분야의 다양한 LLM 응용에 따른 특정 도전과 필요를 해결하기 위한 과업별 지침이 새롭게 추가되었다. 이를 통해 연구 중인 LLM의 기능과 목적에 맞는 맞춤형의 관련성 높은 보고가 가능해진다.

투명성과 공정성에 대한 강조 강화: 새로운 지침은 '투명성'과 '공정성'을 강조하며, 임상모델에 내재될 수 있는 사회적 편향의 인식 및 해결의 중요성을 부각하였다. 체크리스트는 이러한 개념을 전반에 통합하여, 모델 생애주기의 모든 단계에서 편향과 공정성이 고려되도록 하였다.

모듈형(modular) 프레임워크: 새로운 지침은 모듈형 구조로, 각 연구에서 보고되는 연구설계 및 LLM 과업에 따라 요구사항이 달라진다. 이는 모델 개발부터 평가까지 생의학 LLM 연구의 다양한 응용과 접근법을 반영하여, 보다 특화된 보고 항목의 필요성에 대응하기 위함이다.

Table 3. TRIPOD-LLM 초록 항목

섹션(section)	항목(item)	체크리스트 항목(checklist item)	연구설계(Research design)	LLM 과업(LLM task)
제목(Title)	2a	연구가 LLM의 개발, 미세조정, 성능 평가임을 밝히고, 해당 작업, 대상 집단, 예측하고자 하는 결과를 명시할 것	All	All
배경(Background)	2b	의료 맥락, 활용 사례, LLM 성능 개발 또는 평가의 근거를 간단히 설명할 것	E, H	All
목적(Objectives)	2c	연구목적을 명시하되, LLM 개발, 미세조정 및/또는 평가에 관한 것인지 포함할 것	All	All
	2d	연구환경의 주요 요소를 기술할 것	All	All
	2e	연구에 사용된 모든 데이터를 상세히 기술하고, 데이터 분할 및 선택적 사용 여부를 명확히 기술할 것	M, D, E	All
	2f	사용된 LLM의 이름과 버전을 명확히 밝힐 것	All	All
	2g	LLM 구축과정(미세조정, 보상 모델링, RLHF 등 포함)을 간단히 요약할 것	M, D	All
	2h	LLM이 수행한 구체적 작업(예: 의학 QA, 요약, 추출 등)을 기술하고, 최종 LLM의 주요 입력 및 출력을 강조할 것	All	All
방법(Methods)	2i	평가에 사용된 데이터 세트/집단과 평가 엔드포인트를 명시하고, 해당 정보를 학습/튜닝에서 배제했는지를 밝히며, LLM 성능 평가에 사용된 측정방법을 상세히 기술할 것	All	All
결과(Results)	2j	주요 결과의 전반적 보고와 해석을 제공할 것	All	All
논의(Discussion)	2k	결과에 비추어 발생할 수 있는 광범위한 시사점이나 우려 사항을 명확히 기술할 것	All	All
기타(Other)	2l	등록번호와 레지스트리/저장소 이름(해당 시)을 명시할 것	H	All

TRIPOD, transparent reporting of a multivariable model for individual prognosis or diagnosis; LLM, large language model; RLHF, reinforcement learning with human feedback; E, LLM evaluation; H, LLM evaluation in healthcare settings; M, LLM methods; D, *de novo* LLM development; QA, long-form question answering.

ical practice guidelines) [44,45] 개발의 경험에서 영감을 받은 것이다. 지침에 대한 공공 의견은, 접근성을 높이기 위해 여러 경로—프로젝트별 GitHub 저장소, TRIPOD-LLM 웹사이트, 메인 TRIPOD 웹사이트(https://www.tripod-statement.org/)—를 통해 수집할 예정이다. 모호하거나 중복된 표현 등 지침의 가독성과 내용 전반에 대한 의견 모두를 환영한다. 예를 들어, 사용자는 실제 적용 가능성을 높이기 위한 항목 변경, 새로운 항목 추가, 특정 연구설계 또는 LLM 작업 모듈에 배정된 항목 추가/삭제, 연구설계 또는 LLM 작업 모듈 범주 변경 등을 제안할 수 있다.

전문가 패널은 3개월마다 회의를 열어 업데이트를 논의한다. 회의 전, 패널 구성원들은 그동안의 주요 문헌을 검토하여 업데이트에 참고한다. 업데이트 단위는 지침에 명시된 체크리스트 항목, 연구설계 범주, LLM 작업 범주가 된다. 회의에서 패널은 현재 지침의 상태를 점검하고, 공공 의견, 문헌 검토, 주제 전문성을 고려하여 개정사항을 제안한다. 운영위원회는 논의내용을 반영해 지침을 수정하고, 이를 전문가 패널에 회람하여 최종 검토 및 승인을 받는다. 검토 결과에 따라 TRIPOD-LLM 지침의 각 구성요소(항목, 연구설계, LLM 작업)에는 다음과 같은 조치가 취해질 수 있다

[42]: (1) 변경 없음; (2) 실질적 내용의 수정(명확성을 위한 간단한 문구 수정이나 오 · 탈자 교정 등은 수정에 해당하지 않음); (3) 하나 이상의 구성요소를 통합(통합은 동일한 구성요소 유형 내에서만 이루어짐); (4) 하나의 구성요소를 둘 이상으로 분리(분리 역시 동일한 구성요소 유형 내에서만 적용); (5) 해당 구성요소를 지침에서 삭제.

새로운 버전의 지침이 공개될 때마다 TRIPOD-LLM 웹사이트와 메인 TRIPOD 웹사이트(https://www.tripod-statement.org/), EQUATOR Network 웹사이트(https://www.equator-network.org/reporting-guidelines/) 및 소셜 미디어 계정을 통해 즉시 공지된다. 저널 편집자들에게도 이메일로 업데이트 소식이 전달되어, 저자 지침이 항상 최신 버전을 참조하도록 한다. 사용자는 자신이 활용한 지침의 버전을 반드시 인용해야 한다. 각 검토 회의마다 전문가 패널의 전문성, 다양성, 대표성을 점검하며, 필요 시 신규 패널을 위촉한다. 또한 전문가 패널 구성원은 해당 분야에 중대한 변화가 발생하여 긴급한 논의가 필요하다고 판단될 때, 임시(ad hoc) 검토를 요청할 권한도 가진다.

고찰

TRIPOD-LLM은 생물의학 및 의료 분야에서 급속하게 발전하고 있는 LLM 분야에서 연구자와 학술지, 의료전문가, 상업적 · 비상업적 LLM 개발자, 그리고 의료기관을 위한 안내를 목적으로 개발되었다. 이 지침은 LLM의 개발, 튜닝, 평가를 다루는 연구에 대한 최소한의 보고 권고사항을 제시한다. TRIPOD-LLM 항목에 따라 보고함으로써 연구자는 LLM 연구방법의 질을 이해하고 평가할 수 있다. 또한 연구결과의 투명성을 높이고, 연구결과의 과해석을 줄이고, 재현성과 복제를 용이하게 하며, LLM의 실제 적용에도 도움이 된다.

모델 생애주기 전반에 걸친 투명성은 본 가이드라인에서 강하게 강조되는 핵심 요소이다. LLM 생애주기 각 단계에서의 세부적인 문서화가 요구된다[46]. 예를 들어, 개발 및 미세조정 단계에서는 학습 데이터의 출처와 전처리(preprocessing) 과정을 공개하는 것이 중요하다. 또한 LLM의 버전과 기존의 파운데이션 모델에 대한 미세조정 또는 얼라인먼트 과정에 대한 세부 사항을 투명하게 보고해야 LLM 간의 공정한 비교가 가능하다. 여기에는 학습 데이터 수집 시점의 기준일(cut-off date)을 명시하여, 데이터 세트의 시간적 적합성과 평가 시의 데이터 유출 또는 오염 가능성을 분명히 하는 내용이 포함된다. 아울러, 연구에서는 모델 버전의 날짜, 데이터 수집 중 모델이 고정(frozen)되어 있었는지 혹은 동적으로 업데이트 되었는지도 반드시 기록해야 한다. 입력 데이터의 투명성 역시 필수적이다. LLM은 보통 여러 공개 대규모 데이터 세트로 훈련되므로, 사회적 편향이나 불평등(낙인을 찍는 용어, 집단 간 통계적 위험 배분 등)이 내재화될 위험이 있다. 따라서 데이터 소스의 선별과 잠재적 편향에 대한 체계적이고 투명한 접근이 요구된다[20,32,47-50].

TRIPOD-LLM 지침은 인간의 통찰과 감독(human insight and oversight)도 중요한 구성요소로 다루며, 이는 LLM의 책임감 있는 실제 적용을 위해 필수적인 요소이다(다만, 배포 신뢰성과 관찰 가능성 자체는 본 논문의 범위를 벗어난다)[51-53]. 본 가이드라인은 LLM의 예상 적용 맥락에 대한 보고를 강화하고, 해당되는 경우 LLM에 부여된 자율성 수준을 명확히 보고할 것을 요구한다. 또한 데이터 세트 구축과 평가 시의 품질관리 절차(예: 평가자 자격, 이중 평가 요구사항, 평가자에게 제공되는 구체적 지침 등)도 강조하며, 이를 통해 텍스트 평가의 미묘한 부분까지 포착하여 안전성과 성능에 대한 신뢰성 있는 평가가 가능하도록 한다.

프롬프트와 과업 특이적 성과 보고는 LLM의 고유 특성으로 인해 필수적으로 추가된 항목이다. 프롬프트 엔지니어링 방법의 차이는 LLM 성능에 큰 영향을 미칠 수 있고, 이는 벤치마크 비교나 실제 적용 가능성을 왜곡할 위험이 있다[54,55]. 따라서 관련이 있는 경우라면 프롬프트 개발에 사용된 데이터 소스, LLM 모델명과 버전, 수행된 전처리 단계, 프롬프트 엔지니어링 방법을 모두 상세히 기술해야 한다. 이를 통해 프롬프트가 LLM으로부터 안정적이고 재현 가능한 성능을 이끌어내도록 설계되었는지 확인할 수 있다. 추가로, 평가 환경(지침, 인터페이스, 평가에 참여한 집단의 특성 등)에 대한 명확한 보고도 요구된다. 이는 LLM의 성능이 실제 적용 환경과 유사한 조건에서 평가되었는지를 강조하여 그 실용적 유용성을 측정하는 믿을 만한 기준으로 삼기 위함이다.

TRIPOD-LLM의 주요 사용자 및 수혜자는 (1) 논문을 저술하는 학계 및 산업계 연구자, (2) 연구 논문을 평가하는 학술지 편집자 및 심사위원, (3) 투명성과 연구 질 제고의 혜택을 받을 기타 이해관계자(연구 공동체, 학술기관, 정책 입안자, 연구비 지원기관, 규제기관, 환자, 연구 참여자, 산업계, 일반 대중 등)가 될 것이다. 편집자, 출판사, 산업계에서는 학술지 저자 지침 내에 TRIPOD-LLM 링크를 명시하고, 논문 제출 및 심사과정에서 활용을 권고하며, 권고사항 준수를 표준으로 삼을 것을 권장한다. 또한 연구비 지원기관에서도 LLM 연구 지원 신청 시 TRIPOD-LLM에 따라 보고계획을 포함하도록 요구하여, 연구의 낭비를 줄이고 연구비의 효용을 높일 수 있을 것이다.

이 가이드라인은 주로 텍스트 전용 LLM을 염두에 두고 개발되었으나, 최근 LLM이 통합된 멀티모달(multi-modal) 모델(예: 비전-언어 모델[56])도 빠르게 등장하고 있어, 신속하고 유연한 보고 지침이 특히 필요하다. 보고 고려사항의 상당수는 텍스트 LLM과 멀티모달 모델 모두에 적용되지만, 예를 들어 비전-언어 모델에서는 텍스트와 이미지의 전처리 과정을 모두 보고해야 한다. 그러나 향후 버전에는 멀티모달 특유의 고려사항도 반영할 필요가 있다. 예컨대, 영상 데이터를 활용하는 LLM 개발 연구는 영상 획득에 대한 세부 정보를 보고해야 한다. 당분간 LLM을 주된 구성요소로 포

함하는 방법의 개발/평가 연구는 TRIPOD-LLM 지침을 사용할 것을 권고하지만, 이 부분은 해석에 따라 다를 수 있음을 인정한다. 사용자는 재현성과 이해 가능성, 투명성이라는 목표를 염두에 두고, 상식에 기반해 적절한 보고 지침을 선택하고 TRIPOD-LLM 항목 중 멀티모달 LLM에 적용할 수 있는 요소를 해석하여 보고해야 한다. 필요 시 방사선학(radiomics) 등 AI 타 분야의 방법론 가이드라인도 참고할 수 있다[57,58].

임상 AI의 모델 메타데이터를 요약하는 model card의 생성, 검증, 인증, 유지 · 관리에는 Coalition for Health AI [59], Epic AI Labs [60,61]와 같은 검증 연구소나 내부 검증기준이 중요한 역할을 하게 될 것이다. TRIPOD-LLM 기준은 이러한 연구소들이 규제기준에 부합하는 LLM 검증방안을 마련하고(예: 미국 바이든 행정부의 '안전하고 신뢰할 수 있는 인공지능에 관한 행정명령,' 미국 AI Safety Institute [62], 미국 ONC HTI-1 Final Rule [63], EU AI Act [64]), 임상 AI의 신뢰성과 유용성에 대해 환자와 임상의, 그리고 이해관계자의 신뢰를 형성하는 데 기여할 것으로 기대된다.

LLM을 평가하고 검증하려면 특수한 전문성과 자원이 필요하다는 점도 강조할 필요가 있다. LLM을 공정하고 안전하게 도입하려면 개발에 대한 투자뿐 아니라, 대형 학술기관 외의 환경에서도 견고한 검증이 가능하도록 하는 인프라 구축에 대한 투자도 병행되어야 한다. LLM이 시간과 지역에 따라 변화하는 맥락을 내포하고 있으므로 모델의 성능과 공정성이 시점이나 기관에 따라 달라질 수 있다는 점을 고려할 때, 이 체크리스트는 LLM 평가를 위한 지속적인 과정의 일부로 인식되어야 한다. 사용자에 따라 결과가 달라지는 LLM의 특성으로 인해 기존 ML 모델보다 변화 양상이 더욱 예측하기 어려울 수 있으므로, 단일 시점의 검증 결과만으로 보편적 유효성을 주장하기보다는 효과의 경향성과 이질성을 파악하는 데 더 집중해야 한다.

현재 TRIPOD-LLM 체크리스트의 한계는, 이 분야의 전례 없이 빠른 개발 및 출판 속도로 인해 연구 공동체를 위한 신속한 가이드라인 개발이 불가피했다는 점에서 비롯된다. 본 연구에서는 초기 체크리스트 도출을 위해 신속한 델파이 과정을 거쳤으나, 이로 인해 합의와 입력의 폭이 제한될 수 있었음을 인정한다. 이에 따라, 피드백을 신속하게 반영하고 변화하는 방법론에 적응할 수 있도록 생명력 있는(living) 지침방식을 도입하였다. 이러한 특성상 본 논문에 포함된 체크리스트는 시간이 지나면 구 버전이 될 수 있으므로, 사용자는 항상 https://tripod-llm.vercel.app/에서 최신 버전을 확인해야 한다.

결론

TRIPOD-LLM은 연구자가 자신의 연구를 완전하게 보고할 수 있도록 돕고, LLM 개발자와 연구자, 전문가 심사자(peer reviewer), 편집자, 정책 입안자, 최종 사용자(예: 의료전문가), 그리고 환자가 LLM 기반 연구의 데이터, 방법, 결과, 결론을 명확히 이해할 수 있도록 지원하는 것을 목표로 한다. TRIPOD-LLM 보고 권고사항을 준수하면 연구에 소요되는 시간과 노력, 비용을 가장 효율적으로 활용할 수 있으며, LLM 연구의 가치를 높이고 긍정적 영향을 극대화할 수 있을 것이다.

온라인 콘텐츠

연구 방법, 추가 참고문헌, Nature Portfolio 보고 요약, 소스 데이터, 확장 데이터, 보충 정보, 감사의 글, 동료평가 정보, 저자 기여 및 이해관계 관련 상세 내용, 데이터 및 코드 이용 가능성 진술 등은 https://doi.org/10.1038/s41591-024-03425-5에서 확인할 수 있다.

Additional information

[1]*Laboratory for Computational Physiology, Massachusetts Institute of Technology, Cambridge, MA, USA*
[2]*Department of Critical Care, Guy's and St Thomas' NHS Foundation Trust, London, UK*
[3]*Artificial Intelligence in Medicine Program, Mass General Brigham, Harvard Medical School, Boston, MA, USA*
[4]*Department of Medicine, University of Wisconsin—Madison, Madison, WI, USA*
[5]*Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA*
[6]*Tasmanian School of Medicine, College of Health and Medicine, University of Tasmania, Hobart, Tasmania, Australia*
[7]*Department of Population Health, NYU Grossman School of Medicine and Langone Health, New York, NY, USA*
[8]*Department of Radiation Oncology, Brigham and Women's Hospital/Dana–Farber Cancer Institute, Boston, MA, USA*
[9]*USC Institute of Urology and Catherine and Joseph Aresty Department of Urology, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA*
[10]*Artificial Intelligence Center, USC Institute of Urology, University of Southern California, Los Angeles, CA, USA*
[11]*National Library of Medicine, National Institutes of Health, U.S. Department of Health and Human Services, Bethesda, MD, USA*
[12]*Department of Computer Science, Loyola University, Chicago, IL, USA*
[13]*Department of Dermatology, Stanford School of Medicine, Redwood City, CA, USA*
[14]*Department of Biomedical Data Science, Stanford School of Medicine, Redwood City, CA, USA*
[15]*Cognitive Science, Aarhus University, Aarhus, Denmark*
[16]*Mass General Brigham, Boston, MA, USA*
[17]*Faculty of Medicine and Dentistry, University of Alberta, Edmonton, AB, Canada*
[18]*Computational Health Informatics Program, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA*
[19]*Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA*
[20]*Departamento de Telematica, Universidad del Cauca, Popayan, Colombia*
[21]*Dana–Farber Cancer Institute, Boston, MA, USA*

[22]*Department of Radiology, Emory University School of Medicine, Atlanta, GA, USA*
[23]*Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology, and Musculoskeletal Sciences, University of Oxford, Oxford, UK*
[24]*UK EQUATOR Centre, Nuffield Department of Orthopaedics, Rheumatology, and Musculoskeletal Sciences, University of Oxford, Oxford, UK*
[25]*Julius Center for Health Sciences and Primary Care, UMC Utrecht, Utrecht University, Utrecht, the Netherlands*
[26]*Health Innovation Netherlands, Utrecht, the Netherlands*
[27]*Division of Pulmonary, Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA*
[28]*Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA*
[29]*See Authors' contributions*

ORCID

Jack Gallifant: https://orcid.org/0000-0003-1306-2334
Majid Afshar: https://orcid.org/0000-0002-6368-4652
Saleem Ameen: https://orcid.org/0000-0002-2549-4540
Yindalon Aphinyanaphongs: https://orcid.org/0000-0001-8605-5392
Shan Chen: https://orcid.org/0000-0001-7999-7410
Giovanni Cacciamani: https://orcid.org/0000-0002-8892-5539
Dmitriy Dligach: https://orcid.org/0000-0002-2585-2707
Roxana Daneshjou: https://orcid.org/0000-0001-7988-9356
Chrystinne Fernandes: https://orcid.org/0000-0002-8623-9500
Lasse Hyldig Hansen: https://orcid.org/0009-0005-1556-679X
Adam Landman: https://orcid.org/0000-0002-2166-0521
Timothy Miller: https://orcid.org/0000-0003-4513-403X
Amy Moreno: https://orcid.org/0000-0001-6762-6807
David Restrepo: https://orcid.org/0000-0002-3789-1957
Guergana Savova: https://orcid.org/0000-0002-5887-200X
Renato Umeton: https://orcid.org/0000-0002-5561-6932
Judy Wawira Gichoya: https://orcid.org/0000-0002-1097-316X
Gary S. Collins: https://orcid.org/0000-0002-2772-2316
Karel G. M. Moons: https://orcid.org/0000-0003-2118-004X
Leo A. Celi: https://orcid.org/0000-0003-2118-004X
Danielle S. Bitterman: https://orcid.org/0000-0003-0345-2232

Authors' contributions

DSB, JG, LAC, GSC, and KGMM were on the steering group that directed the guideline-development process. SC, CF, DR, GS, TM, DDF, RU, LHH, YA, JWG, LGM, NM, and RD were members of the expert panel. DSB and JG drafted the initial list of candidate items.

These authors contributed equally: Majid Afshar, Saleem Ameen, Yindalon Aphinyanaphongs, Shan Chen, Giovanni Cacciamani, Dina Demner-Fushman, Dmitriy Dligach, Roxana Daneshjou, Chrystinne Fernandes, Lasse Hyldig Hansen, Adam Landman, Lisa Lehmann, Liam G. McCoy, Timothy Miller, Amy Moreno, Nikolaj Munch, David Restrepo, Guergana Savova, Renato Umeton, and Judy Wawira Gichoya.

Conflict of interest

DSB is an associate editor at *Radiation Oncology* and HemOnc.org, receives research funding from the American Association for Cancer Research, and provides advisory and consulting services for Mercurial AI. DDF is an associate editor at the *Journal of the American Medical Informatics Association*, is a member of the editorial board of *Scientific Data*, and receives funding from the intramural research program at the US National Library of Medicine, NIH. JWG is a member of the editorial board of *Radiology: Artificial Intelligence*, *BJR Artificial Intelligence*, and *NEJM AI*. All other authors declare no potential conflict of interest relevant to this article.

Funding

JG is funded by the US National Institutes of Health (NIH) (U54 TW012043-01 and OT2OD032701). SC is supported by the NIH (U54CA274516-01A1). GSC is supported by Cancer Research UK (program grant C49297/A27294) and by the Engineering and Physical Sciences Research Council grant for 'Artificial intelligence innovation to accelerate health research' (EP/Y018516/1), and is a National Institute for Health and Care Research (NIHR) senior investigator. The views expressed in this article are those of the author and not necessarily those of the NIHR or the Department of Health and Social Care. YA is partially supported by the NIH (3UL1TR001445-05) and the National Science Foundation Award (1928614 and 2129076). LAC is funded by the NIH (U54 TW012043-01, OT2OD032701 and R01EB017205). DSB is supported by the NIH (U54CA274516-01A1 and R01CA294033-01) and the American Cancer Society and American Society for Radiation Oncology grant ASTRO-CSDG-24-1244514-01-CTPS. JWG receives support from 2022 Robert Wood Johnson Foundation Harold Amos Medical Faculty Development Program, a Radiological Society of North America Health Disparities grant (EIHD2204), the Lacuna Fund (67), the Gordon and Betty Moore Foundation, an NIH (National Institute of Biomedical Imaging and Bioengineering) MIDRC grant under contracts 75N92020C00008 and 75N92020C00021, and an NHLBI award (R01HL167811).

Data availability

Research data are available at https://doi.org/10.1038/s41591-024-03425-5

Acknowledgments

None.

Supplementary materials

The online version contains supplementary material available at https://doi.org/10.1038/s41591-024-03425-5

Supplement 1. Completed TRIPOD-LLM checklist for NYUTron.

Supplement 2. Fillable TRIPOD-LLM checklist.

Supplement 3. TRIPOD-LLM expanded checklist (explanation and elaboration light).

References

1. Chen Z, Cano AH, Romanou A, Bonnet A, Matoba K, Salvi F, Pagliardini M, Fan S, Kopf A, Mohtashami A, Sallinen A. Meditron-70B: scaling medical pretraining for large language models. arXiv [Preprint] 2023 Nov 27. https://doi.org/10.48550/arXiv.2311.16079
2. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, Almeida D, Altenschmidt J, Altman S, Anadkat S, Avila R, Babuschkin I, Balaji S, Balcom V, Baltescu P, Bao H, Bavarian M, Belgum J, Bello I, Berdine J, Bernadett-Shapiro G, Berner C, Bogdonoff L, Boiko O, Boyd M, Brakman AL, Brockman G, Brooks T, Brundage M, Button K, Cai T, Campbell R, Cann A, Carey B, Carlson C, Carmichael R, Chan B, Chang C, Chantzis F, Chen D, Chen S, Chen R, Chen J, Chen M, Chess B, Cho C, Chu C, Chung HW, Cummings D, Currier J, Dai Y, Decareaux C, Degry T, Deutsch N, Deville D, Dhar A, Dohan D, Dowling S, Dunning S, Ecoffet A, Eleti A, Eloundou T, Farhi D, Fedus L, Felix N, Fishman SP, Forte J, Fulford I, Gao L, Georges E, Gibson C, Goel V, Gogineni T, Goh G, Gontijo-Lopes R, Gordon J, Grafstein M, Gray S, Greene R, Gross J, Gu SS, Guo Y, Hallacy C, Han J, Harris J, He Y, Heaton M, Heidecke J, Hesse C, Hickey A, Hickey W, Hoeschele P, Houghton B, Hsu K, Hu S, Hu X, Huizinga J, Jain S, Jain S. GPT-4 technical report. arXiv [Preprint] 2023 Dec 19. https://doi.org/10.48550/arXiv.2303.08774
3. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, Scales N, Tanwani A, Cole-Lewis H, Pfohl S, Payne P, Seneviratne M, Gamble P, Kelly C, Babiker A, Scharli N, Chowdhery A, Mansfield P, Demner-Fushman D, Aguera Y Arcas B, Webster D, Corrado GS, Matias Y, Chou K, Gottweis J, Tomasev N, Liu Y, Rajkomar A, Barral J, Semturs C, Karthikesalingam A, Natarajan V. Large language models encode clinical knowledge. Nature 2023;620:172-180. https://doi.org/10.1038/s41586-023-06291-2
4. Tai-Seale M, Baxter SL, Vaida F, Walker A, Sitapati AM, Osborne C, Diaz J, Desai N, Webb S, Polston G, Helsten T, Gross E, Thackaberry J, Mandvi A, Lillie D, Li S, Gin G, Achar S, Hofflich H, Sharp C, Millen M, Longhurst CA. AI-generated draft replies integrated into health records and physicians' electronic communication. JAMA Netw Open 2024;7:e246565. https://doi.org/10.1001/jamanetworkopen.2024.6565
5. Tierney AA, Gayre G, Hoberman B, Mattern B, Ballesca M, Kipnis P, Liu V, Lee K. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. NEJM Catal Innov Care Deliv 2024;5:CAT.23.0404. https://doi.org/10.1056/CAT.23.0404
6. Jiang LY, Liu XC, Nejatian NP, Nasir-Moin M, Wang D, Abidin A, Eaton K, Riina HA, Laufer I, Punjabi P, Miceli M, Kim NC, Orillac C, Schnurman Z, Livia C, Weiss H, Kurland D, Neifert S, Dastagirzada Y, Kondziolka D, Cheung AT, Yang G, Cao M, Flores M, Costa AB, Aphinyanaphongs Y, Cho K, Oermann EK. Health system-scale language models are all-purpose prediction engines. Nature 2023;619:357-362. https://doi.org/10.1038/s41586-023-06160-y
7. Cohen MK, Kolt N, Bengio Y, Hadfield GK, Russell S. Regulating advanced artificial agents. Science 2024;384:36-38. https://doi.org/10.1126/science.adl0625
8. Mesko B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med 2023;6:120. https://doi.org/10.1038/s41746-023-00873-0
9. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ 2015;350:g7594. https://doi.org/10.1136/bmj.g7594
10. EQUATOR Network. Reporting guidelines [Internet]. EQUATOR Network [cited 2024 Jun 1]. Available from: https://www.equator-network.org/reporting-guidelines/

11. Collins GS, Moons KG, Dhiman P, Riley RD, Beam AL, Van Calster B, Ghassemi M, Liu X, Reitsma JB, van Smeden M, Boulesteix AL, Camaradou JC, Celi LA, Denaxas S, Denniston AK, Glocker B, Golub RM, Harvey H, Heinze G, Hoffman MM, Kengne AP, Lam E, Lee N, Loder EW, Maier-Hein L, Mateen BA, McCradden MD, Oakden-Rayner L, Ordish J, Parnell R, Rose S, Singh K, Wynants L, Logullo P. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024; 385:e078378. https://doi.org/10.1136/bmj-2023-078378
12. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med 2020;26:1364-1374. https://doi.org/10.1038/s41591-020-1034-x
13. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, Denniston AK, Faes L, Geerts B, Ibrahim M, Liu X, Mateen BA, Mathur P, McCradden MD, Morgan L, Ordish J, Rogers C, Saria S, Ting DS, Watkinson P, Weber W, Wheatstone P, McCulloch P. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med 2022;28:924-933. https://doi.org/10.1038/s41591-022-01772-9
14. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, Arnaout R, Kohane IS, Saria S, Topol E, Obermeyer Z, Yu B, Butte AJ. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med 2020;26:1320-1324. https://doi.org/10.1038/s41591-020-1041-y
15. Agrawal M, Hegselmann S, Lang H, Kim Y, Sontag D. Large language models are few-shot clinical information extractors. In: Goldberg Y, Kozareva Z, Zhang Y, editors. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing; 2022 Dec; Abu Dhabi, United Arab Emirates. Association for Computational Linguistics; 2022. p. 1998-2022. https://doi.org/10.18653/v1/2022.emnlp-main.130
16. Liu X, McDuff D, Kovacs G, Galatzer-Levy I, Sunshine J, Zhan J, Poh MZ, Liao S, Di Achille P, Patel S. Large language models are few-shot health learners. arXiv [Preprint] 2023 May 24. https://doi.org/10.48550/arXiv.2305.15525
17. Salazar J, Liang D, Nguyen TQ, Kirchhoff K. Masked language model scoring. In: Jurafsky D, Chai J, Schluter N, Tetreault J, editors. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020 Jul; Online. Association for Computational Linguistics; 2022. p. 2699-2712. https://doi.org/10.18653/v1/2020.acl-main.240
18. Wang A, Singh A, Michael J, Hill F, Levy O, Bowman S. GLUE: a multi-task benchmark and analysis platform for natural language understanding. In: Linzen T, Chrupala G, Alishahi A, editors. Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP; 2018 Nov; Brussels, Belgium. Association for Computational Linguistics; 2018. p. 353-355. https://doi.org/10.18653/v1/W18-5446
19. Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: a method for automatic evaluation of machine translation. In: Isabelle P, Charniak E, Lin D, editors. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics; 2002 Jul; Philadelphia, USA. Association for Computational Linguistics; 2002. p. 311-318. https://doi.org/10.3115/1073083.1073135
20. Zack T, Lehman E, Suzgun M, Rodriguez JA, Celi LA, Gichoya J, Jurafsky D, Szolovits P, Bates DW, Abdulnour RE, Butte AJ, Alsentzer E. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. Lancet Digit Health 2024;6:e12-e22. https://doi.org/10.1016/S2589-7500(23)00225-X
21. Chen S, Guevara M, Moningi S, Hoebers F, Elhalawani H, Kann BH, Chipidza FE, Leeman J, Aerts HJWL, Miller T, Savova GK, Gallifant J, Celi LA, Mak RH, Lustberg M, Afshar M, Bitterman DS. The effect of using a large language model to respond to patient messages. Lancet Digit Health 2024;6: e379-e381. https://doi.org/10.1016/S2589-7500(24)00060-8
22. Peng C, Yang X, Chen A, Smith KE, PourNejatian N, Costa AB, Martin C, Flores MG, Zhang Y, Magoc T, Lipori G, Mitchell DA, Ospina NS, Ahmed MM, Hogan WR, Shenkman EA, Guo Y, Bian J, Wu Y. A study of generative large language model for medical research and healthcare. NPJ Digit Med 2023;6:210. https://doi.org/10.1038/s41746-023-00958-w
23. Zakka C, Shad R, Chaurasia A, Dalal AR, Kim JL, Moor M, Fong R, Phillips C, Alexander K, Ashley E, Boyd J, Boyd K, Hirsch K, Langlotz C, Lee R, Melia J, Nelson J, Sallam K, Tullis S, Vogelsong MA, Cunningham JP, Hiesinger W. Almanac: retrieval-augmented language models for clinical medicine. NEJM AI 2024;1:10.1056/aioa2300068. https://doi.org/10.1056/aioa2300068
24. Keloth VK, Hu Y, Xie Q, Peng X, Wang Y, Zheng A, Selek M, Raja K, Wei CH, Jin Q, Lu Z, Chen Q, Xu H. Advancing entity recognition in biomedicine via instruction tuning of large lan-

guage models. Bioinformatics 2024;40:btae163. https://doi.org/10.1093/bioinformatics/btae163

25. Guevara M, Chen S, Thomas S, Chaunzwa TL, Franco I, Kann BH, Moningi S, Qian JM, Goldstein M, Harper S, Aerts HJ, Catalano PJ, Savova GK, Mak RH, Bitterman DS. Large language models to identify social determinants of health in electronic health records. NPJ Digit Med 2024;7:6. https://doi.org/10.1038/s41746-023-00970-0
26. Jin Q, Yang Y, Chen Q, Lu Z. GeneGPT: augmenting large language models with domain tools for improved access to biomedical information. Bioinformatics 2024;40:btae075. https://doi.org/10.1093/bioinformatics/btae075
27. Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman H, Cool JA, Kanjee Z, Parsons AS, Ahuja N, Horvitz E, Yang D, Milstein A, Olson AP, Rodman A, Chen JH. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open 2024;7:e2440969. https://doi.org/10.1001/jamanetworkopen.2024.40969
28. Zaretsky J, Kim JM, Baskharoun S, Zhao Y, Austrian J, Aphinyanaphongs Y, Gupta R, Blecker SB, Feldman J. Generative artificial intelligence to transform inpatient discharge summaries to patient-friendly language and format. JAMA Netw Open 2024;7:e240357. https://doi.org/10.1001/jamanetworkopen.2024.0357
29. Han L, Erofeev G, Sorokina I, Gladkoff S, Nenadic G. Examining large pre-trained language models for machine translation: what you don't know about it. Proceedings of the Seventh Conference on Machine Translation (WMT); 2022 Dec; Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics; 2022. p. 908-919.
30. Yoon W, Chen S, Gao Y, Zhao Z, Dligach D, Bitterman DS, Afshar M, Miller T. LCD benchmark: long clinical document benchmark on mortality prediction for language models. J Am Med Inform Assoc 2025;32:285-295. https://doi.org/10.1093/jamia/ocae287
31. Goodman KE, Yi PH, Morgan DJ. AI-generated clinical summaries require more than accuracy. JAMA 2024;331:637-638. https://doi.org/10.1001/jama.2024.0555
32. Gallifant J, Fiske A, Levites Strekalova YA, Osorio-Valencia JS, Parke R, Mwavu R, Martinez N, Gichoya JW, Ghassemi M, Demner-Fushman D, McCoy LG, Celi LA, Pierce R. Peer review of GPT-4 technical report and systems card. PLOS Digit Health 2024;3:e0000417. https://doi.org/10.1371/journal.pdig.0000417
33. Wornow M, Xu Y, Thapa R, Patel B, Steinberg E, Fleming S, Pfeffer MA, Fries J, Shah NH. The shaky foundations of large language models and foundation models for electronic health records. NPJ Digit Med 2023;6:135. https://doi.org/10.1038/s41746-023-00879-8
34. Chang CT, Farah H, Gui H, Rezaei SJ, Bou-Khalil C, Park YJ, Swaminathan A, Omiye JA, Kolluri A, Chaurasia A, Lozano A, Heiman A, Jia AS, Kaushal A, Jia A, Iacovelli A, Yang A, Salles A, Singhal A, Narasimhan B, Belai B, Jacobson BH, Li B, Poe CH, Sanghera C, Zheng C, Messer C, Kettud DV, Pandya D, Kaur D, Hla D, Dindoust D, Moehrle D, Ross D, Chou E, Lin E, Haredasht FN, Cheng G, Gao I, Chang J, Silberg J, Fries JA, Xu J, Jamison J, Tamaresis JS, Chen JH, Lazaro J, Banda JM, Lee JJ, Matthys KE, Steffner KR, Tian L, Pegolotti L, Srinivasan M, Manimaran M, Schwede M, Zhang M, Nguyen M, Fathzadeh M, Zhao Q, Bajra R, Khurana R, Azam R, Bartlett R, Truong ST, Fleming SL, Raj S, Behr S, Onyeka S, Muppidi S, Bandali T, Eulalio TY, Chen W, Zhou X, Ding Y, Cui Y, Tan Y, Liu Y, Shah NH, Daneshjou R. Red teaming large language models in medicine: real-world insights on model behavior. medRxiv [Preprint] 2024 Apr 7. https://doi.org/10.1101/2024.04.05.24305411
35. Gallifant J, Chen S, Moreira PJ, Munch N, Gao M, Pond J, Celi LA, Aerts H, Hartvigsen T, Bitterman D. Language models are surprisingly fragile to drug names in biomedical benchmarks. In: Al-Onaizan Y, Bansal M, Chen YN, editors. Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024; 2024 Nov; Miami, USA. Association for Computational Linguistics; 2024. p. 12448-12465. https://doi.org/10.18653/v1/2024.findings-emnlp.726
36. Boyd E. Microsoft and Epic expand AI collaboration to accelerate generative AI's impact in healthcare, addressing the industry's most pressing needs [Internet]. Microsoft; 2023 [cited 2024 Jun 1]. Available from: https://blogs.microsoft.com/blog/2023/08/22/microsoft-and-epic-expand-ai-collaboration-to-accelerate-generative-ais-impact-in-healthcare-addressing-the-industrys-most-pressing-needs/
37. Moreno AC, Bitterman DS. Toward Clinical-Grade Evaluation of Large Language Models. Int J Radiat Oncol Biol Phys 2024;118:916-920. https://doi.org/10.1016/j.ijrobp.2023.11.012
38. Welch Medical Library. Evidence based medicine: evidence grading & reporting [Internet]. Johns Hopkins University [cited 2024 Jun 1]. Available from: https://browse.welch.jhmi.edu/

EBM/EBM_EvidenceGrading

39. Guyatt GH, Oxman AD, Kunz R, Vist GE, Falck-Ytter Y, Schunemann HJ; GRADE Working Group. What is "quality of evidence" and why is it important to clinicians?. BMJ 2008;336: 995-998. https://doi.org/10.1136/bmj.39490.551019.BE
40. Huo B, Cacciamani GE, Collins GS, McKechnie T, Lee Y, Guyatt G. Reporting standards for the use of large language model-linked chatbots for health advice. Nat Med 2023;29: 2988. https://doi.org/10.1038/s41591-023-02656-2
41. Cacciamani GE, Collins GS, Gill IS. ChatGPT: standard reporting guidelines for responsible use. Nature 2023;618:238. https://doi.org/10.1038/d41586-023-01853-w
42. El Mikati IK, Khabsa J, Harb T, Khamis M, Agarwal A, Pardo-Hernandez H, Farran S, Khamis AM, El Zein O, El-Khoury R, Schünemann HJ, Akl EA, Alonso-Coello P, Alper BS, Amer YS, Arayssi T, Barker JM, Bouakl I, Boutron I, Brignardello-Petersen R, Carandang K, Chang S, Chen Y, Cuker A, El-Jardali F, Florez I, Ford N, Grove J, Guyatt GH, Hazlewood GS, Kredo T, Lamontagne F, Langendam MW, Lewin S, Macdonald H, McFarlane E, Meerpohl J, Munn Z, Murad MH, Mustafa RA, Neumann I, Nieuwlaat R, Nowak A, Pardo JP, Qaseem A, Rada G, Righini M, Rochwerg B, Rojas-Reyes MX, Siegal D, Siemieniuk R, Singh JA, Skoetz N, Sultan S, Synnot A, Tugwell P, Turner A, Turner T, Venkatachalam S, Welch V, Wiercioch W. A framework for the development of living practice guidelines in health care. Ann Intern Med 2022;175:1154-1160. https://doi.org/10.7326/M22-0514
43. Cochrane Community. Living systematic reviews [Internet]. Cochrane [cited 2024 Jun 1]. Available from: https://community.cochrane.org/review-development/resources/living-systematic-reviews
44. Akl EA, Meerpohl JJ, Elliott J, Kahale LA, Schünemann HJ. Living systematic reviews: 4. Living guideline recommendations. J Clin Epidemiol 2017;91:47-53. https://doi.org/10.1016/j.jclinepi.2017.08.009
45. Fraile Navarro D, Cheyne S, Hill K, McFarlane E, Morgan RL, Murad MH, Mustafa RA, Sultan S, Tunnicliffe DJ, Vogel JP, White H, Turner T. Methods for living guidelines: early guidance based on practical experience. Paper 5: decisions on methods for evidence synthesis and recommendation development for living guidelines. J Clin Epidemiol 2023;155:118-128. https://doi.org/10.1016/j.jclinepi.2022.12.022
46. Bedoya AD, Economou-Zavlanos NJ, Goldstein BA, Young A, Jelovsek JE, O'Brien C, Parrish AB, Elengold S, Lytle K, Balu S, Huang E, Poon EG, Pencina MJ. A framework for the oversight and local deployment of safe and high-quality prediction models. J Am Med Inform Assoc 2022;29:1631-1636. https://doi.org/10.1093/jamia/ocac078
47. Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large language models in medicine: the potentials and pitfalls : a narrative review. Ann Intern Med 2024;177:210-220. https://doi.org/10.7326/M23-2772
48. Chen S, Gallifant J, Gao M, Moreira P, Munch N, Muthukkumar A, Rajan A, Kolluri J, Fiske A, Hastings J, Aerts H, Anthony B, Celi LA, La Cava WG, Bitterman DS. Cross-care: assessing the healthcare implications of pre-training data on language model bias. Adv Neural Inf Process Syst 2024;37:23756-23795.
49. Hansen LH, Andersen N, Gallifant J, McCoy LG, Stone JK, Izath N, Aguirre-Jerez M, Bitterman DS, Gichoya J, Celi LA. Seeds of stereotypes: a large-scale textual analysis of race and gender associations with diseases in online sources. arXiv [Preprint] 2024 May 8. https://doi.org/10.48550/arXiv.2405.05049
50. Biderman S, Schoelkopf H, Anthony QG, Bradley H, O'Brien K, Hallahan E, Khan MA, Purohit S, Prashanth US, Raff E, Skowron A, Sutawika L, Van Der Wal O. Pythia: a suite for analyzing large language models across training and scaling. Proceedings of the 40th International Conference on Machine Learning; 2023 Jul 23-29; Honolulu, USA. PMLR; 2023.
51. Bowman SR, Hyun J, Perez E, Chen E, Pettit C, Heiner S, Lukosiute K, Askell A, Jones A, Chen A, Goldie A, Mirhoseini A, McKinnon C, Olah C, Amodei D, Amodei D, Drain D, Li D, Tran-Johnson E, Kernion J, Kerr J, Mueller J, Ladish J, Landau J, Ndousse K, Lovitt L, Elhage N, Schiefer N, Joseph N, Mercado N, DasSarma N, Larson R, McCandlish S, Kundu S, Scott Johnston, Kravec S, El Showk S, Fort S, Telleen-Lawton T, Brown T, Henighan T, Hume T, Bai Y, Hatfield-Dodds Z, Mann B, Kaplan J. Measuring progress on scalable oversight for large language models. arXiv [Preprint] 2022 Nov 11. https://doi.org/10.48550/arXiv.2211.03540
52. McAleese N, Pokorny RM, Uribe JF, Nitishinskaya E, Trebacz M, Leike J. LLM critics help catch LLM bugs. arXiv [Preprint] 2024 Jun 28. https://doi.org/10.48550/arXiv.2407.00215
53. Burns C, Izmailov P, Kirchner JH, Baker B, Gao L, Aschenbrenner L, Chen Y, Ecoffet A, Joglekar M, Leike J, Sutskever I. Weak-to-strong generalization: eliciting strong capabilities with weak supervision. arXiv [Preprint] 2023 Dec 14. https://doi.org/10.48550/arXiv.2312.09390

54. Chen S, Li Y, Lu S, Van H, Aerts HJ, Savova GK, Bitterman DS. Evaluating the ChatGPT family of models for biomedical reasoning and classification. J Am Med Inform Assoc 2024;31:940-948. https://doi.org/10.1093/jamia/ocad256
55. Chen S, Kann BH, Foote MB, Aerts HJ, Savova GK, Mak RH, Bitterman DS. Use of artificial intelligence chatbots for cancer treatment information. JAMA Oncol 2023;9:1459-1462. https://doi.org/10.1001/jamaoncol.2023.2954
56. Lu MY, Chen B, Williamson DF, Chen RJ, Liang I, Ding T, Jaume G, Odintsov I, Le LP, Gerber G, Parwani AV, Zhang A, Mahmood F. A visual-language foundation model for computational pathology. Nat Med 2024;30:863-874. https://doi.org/10.1038/s41591-024-02856-4
57. Kocak B, Akinci D'Antonoli T, Mercaldo N, Alberich-Bayarri A, Baessler B, Ambrosini I, Andreychenko AE, Bakas S, Beets-Tan RG, Bressem K, Buvat I, Cannella R, Cappellini LA, Cavallo AU, Chepelev LL, Chu LC, Demircioglu A, deSouza NM, Dietzel M, Fanni SC, Fedorov A, Fournier LS, Giannini V, Girometti R, Groot Lipman KB, Kalarakis G, Kelly BS, Klontzas ME, Koh DM, Kotter E, Lee HY, Maas M, Marti-Bonmati L, Muller H, Obuchowski N, Orlhac F, Papanikolaou N, Petrash E, Pfaehler E, Pinto Dos Santos D, Ponsiglione A, Sabater S, Sardanelli F, Seebock P, Sijtsema NM, Stanzione A, Traverso A, Ugga L, Vallieres M, van Dijk LV, van Griethuysen JJ, van Hamersvelt RW, van Ooijen P, Vernuccio F, Wang A, Williams S, Witowski J, Zhang Z, Zwanenburg A, Cuocolo R. METhodological RadiomICs Score (METRICS): a quality scoring tool for radiomics research endorsed by EuSoMII. Insights Imaging 2024;15:8. https://doi.org/10.1186/s13244-023-01572-w
58. Lambin P, Leijenaar RT, Deist TM, Peerlings J, de Jong EE, van Timmeren J, Sanduleanu S, Larue RT, Even AJ, Jochems A, van Wijk Y, Woodruff H, van Soest J, Lustberg T, Roelofs E, van Elmpt W, Dekker A, Mottaghy FM, Wildberger JE, Walsh S. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 2017;14:749-762. https://doi.org/10.1038/nrclinonc.2017.141
59. Shah NH, Halamka JD, Saria S, Pencina M, Tazbaz T, Tripathi M, Callahan A, Hildahl H, Anderson B. A nationwide network of health ai assurance laboratories. JAMA 2024;331:245-249. https://doi.org/10.1001/jama.2023.26930
60. Diaz N. Epic releases AI validation suite [Internet]. Becker's Hospital Review; 2024 [cited 2024 May 23]. Available from: https://www.beckershospitalreview.com/ehrs/epic-releases-ai-validation-suite.html
61. Epic-open-source/seismometer [Internet]. GitHub; 2024 [cited 2024 May 23]. Available from: https://github.com/epic-open-source/seismometer
62. National Institute of Standards and Technology (NIST). U.S. Artificial Intelligence Safety Institute [Internet]. NIST; 2023 [cited 2024 May 23]. Available from: https://www.nist.gov/aisi
63. Federal Register. Health data, technology, and interoperability: certification program updates, algorithm transparency, and information sharing [Internet]. Federal Register; 2024 [cited 2024 May 23]. Available from: https://www.federalregister.gov/documents/2024/01/09/2023-28857/health-data-technology-and-interoperability-certification-program-updates-algorithm-transparency-and
64. EU Artificial Intelligence Act. The AI Act Explorer [Internet]. EU Artificial Intelligence Act; 2024 [cited 2024 May 23]. Available from: [cited 2024 May 23]. Available from: https://artificialintelligenceact.eu/ai-act-explorer/

eISSN: 2234-2591
Open Access

Guidelines

Ewha Med J 2025;48(3):e50
https://doi.org/10.12771/emj.2025.00409

The Ewha Medical Journal

CONSORT 2025 statement: 무작위 배정 임상시험 보고 가이드라인 개정판

CONSORT 2025 statement: updated guideline for reporting randomized trials: a Korean translation

Sally Hopewell[1*], An-Wen Chan[2], Gary S. Collins[3], Asbjørn Hróbjartsson[4,5], David Moher[6], Kenneth F. Schulz[7], Ruth Tunn[1], Rakesh Aggarwal[8], Michael Berkwits[9], Jesse A. Berlin[10,11], Nita Bhandari[12], Nancy J. Butcher[13,14], Marion K. Campbell[15], Runcie C. W. Chidebe[16,17], Diana Elbourne[18], Andrew Farmer[19], Dean A. Fergusson[20], Robert M. Golub[21], Steven N. Goodman[22], Tammy C. Hoffmann[23], John P. A. Ioannidis[24], Brennan C. Kahan[25], Rachel L. Knowles[26], Sarah E. Lamb[27], Steff Lewis[28], Elizabeth Loder[29,30], Martin Offringa[13], Philippe Ravaud[31], Dawn P. Richards[32], Frank W. Rockhold[33], David L. Schriger[34], Nandi L. Siegried[35], Sophie Staniszewska[36], Rod S. Taylor[37], Lehana Thabane[38,39], David Torgerson[40], Sunita Vohra[41], Ian R. White[25], Isabelle Boutron[42]

For further information on the authors' affiliations, see ***Additional information***.

배경: 잘 설계되고 적절하게 실행된 무작위 배정 임상시험은 의료 개입의 효과에 대한 가장 신뢰할 수 있는 증거라고 할 수 있다. 그러나 연구 보고의 질이 미흡하다는 많은 증거가 있다. CONSORT(통합 임상시험 보고 기준)는 이러한 보고의 질을 개선하기 위해 고안되었으며, 무작위 임상시험 보고서가 포함해야 할 최소한의 항목을 제시한다. CONSORT는 1996년에 처음 발표된 후 2001년과 2010년에 개정되었다. 여기에서는 최근의 방법론적 발전과 최종 사용자의 피드백을 반영하여 개정된 CONSORT 2025 statement를 소개한다.
방법: 문헌에 대해 범위를 검토하고 CONSORT와 관련된 경험과 이론에 바탕을 둔 프로젝트별 데이터베이스를 개발하여 점검표의 잠재적 변경 목록을 생성하였다. 이 목록은 기존 CONSORT 확장판(유해성, 결과, 비약물적 치료)의 주 저자들과, 관련 보고 지침들(TIDieR), 그리고 개인적인 의견 교환을 포함한 기타 출처에서 제공된 권고사항으로 보강되었다. 점검표의 잠재적 변경 목록은 317명이 참여한 대규모 국제 온라인 3 라운드 델파이 설문조사를 통해 평가되었으며, 초청된 30명의 국제 전문가가 참여한 이틀 간의 온라인 전문가 합의 회의에서 논의되었다.
결과: CONSORT 점검표를 대폭 변경하였다. 새로운 항목 7개를 추가하고, 3개 항목을 수정하고, 1개 항목은 삭제했으며, 주요 CONSORT 확장판의 여러 항목을 통합하였다. 또한 오픈 사이언스에 대한 새로운 섹션을 추가하여 CONSORT 점검표를 재구성하였다. CONSORT 2025 statement는 무작위 임상시험 결과를 보고할 때 반드시 포함해야 하는 30개 항목의 점검표와 임상시험 참여자의 흐름을 문서화하기 위한 다이어그램으로 구성되어 있다. 또한 각 항목의 핵심 요소를 도출하여 글머리 기호 형식으로 정리한 확장형 점검표도 개발하여 CONSORT 2025를 용이하게 이행할 수 있도록 하였다.
결론: 저자, 편집인, 심사자 및 기타 잠재적 사용자는 무작위 임상시험의 원고를 작성하고 평가할 때 CONSORT 2025를 사용하여 임상시험 보고서를 명확하고 투명하게 작성하도록 해야 한다.

*Corresponding email: sally.hopewell@csm.ox.ac.uk
Received: May 7, 2025 Accepted: June 17, 2025
It is a Korean translation of the Hopewell S. et al. CONSORT 2025 statement: updated guideline for reporting randomized trials. Nat Med 2025 Jun;31(6):1776-1783. https://doi.org/10.1038/s41591-025-03635-5.
The translation was done with the permission of the SPIRIT–CONSORT group.
Korean medical terminology is based on the English-Korean Medical Terminology 6th edition, available at: https://term.kma.org/index.asp.
Korean translation was done by Sun Huh (https://orcid.org/0000-0002-8559-8640), Hallym University. The Korean proofreading was conducted by Yoon Joo Seo (https://orcid.org/0000-0002-0202-8352), InfoLumi, and the back-translation was performed by Jeong-Ju Yoo (https://orcid.org/0000-0002-7802-0381), Soonchunhyang University Bucheon Hospital. The back-translation was confirmed by Sally Hopewell (https://orcid.org/0000-0002-6881-6984), the first author of the original CONSORT 2025 statement.

저자 요약

- 무작위 임상시험을 정확하게 해석할 수 있도록 독자에게 임상시험의 방법과 결과에 대한 완전하고 투명한 정보를 제공해야 한다.
- CONSORT 2025 statement는 방법론적 발전과 최종 사용자의 피드백을 반영하여 무작위 배정 임상시험 결과 보고에 대한 개정된 지침을 제공한다.
- CONSORT 2025 statement는 30개 항목으로 구성된 필수 항목 점검표, 임상시험 참여자의 흐름을 문서화하기 위한 다이어그램, 각 점검표 항목의 핵심 요소를 자세히 설명하는 확장된 점검표로 구성되어 있다.
- 저자, 편집인, 검토자 및 기타 잠재적 사용자는 무작위 배정 임상시험 원고를 작성하고 평가할 때 CONSORT 2025를 사용하여 임상시험 보고서가 명확하고 투명하게 작성되도록 해야 한다.

서론

"독자가 연구과정에 무슨 일이 있었는지 유추해야 할 필요가 없도록 명확하게 알려야 한다." Douglas G. Altman [1]

무작위 배정 임상시험은 적절하게 설계, 수행, 분석, 보고되었을 때 일반적으로 의료 중재를 평가하는 데 있어 가장 높은 수준의 근거로 간주된다. 무작위 배정 임상시험의 질에 대한 비판적 평가는 임상시험의 설계, 수행, 분석 및 결과를 철저하고 정확하게 보고할 때만 가능하다. 독자가 임상시험을 정확하게 해석하려면 임상시험의 방법과 결과에 대한 완전하고 투명한 정보가 필요하다. 그러나 무작위 배정 임상시험 보고의 완전성이 불충분하며[2,3] 이러한 불완전한 보고는 중재효과의 추정치에 편향을 초래할 수 있다는 광범위한 증거가 있다[4]. 마찬가지로, 명확하고 투명한 임상시험 프로토콜을 갖추는 것은 중요하다. 일차 결과와 같이 임상시험에 사용될 방법을 미리 지정함으로써, 미공개 사후 변경이 이루어질 가능성을 줄이기 때문이다[5].

무작위 임상시험 보고를 개선하려는 노력은 1990년대 초에 탄력을 받아, 1994년 표준화된 임상시험 보고(Standardised Reporting of Trials, SORT) 및 아실로마르 이니셔티브(Asilomar initiatives)가 탄생하였다. 이러한 이니셔티브는 1996년 CONSORT(임상시험 보고 통합 기준) statement 발표로 이어졌고[6], 2001년에는 설명 및 상세 문서와 함께 개정되었다[7,8]. 이후 2010년에 CONSORT가 개정되었고[9], 이에 대한 설명 및 상세 문서도 함께 개정되었다[10]. 마찬가지로 임상시험 프로토콜에서 완전하고 투명한 보고가 부족하다는 문제로 인해 SPIRIT(표준 프로토콜 항목: 중재적 임상시험을 위한 권고사항) statement가 개발되어 2013년에 발표되었으며[11], 이 statement의 기본 원칙에 대한 설명 및 상세 문서도 함께 발표되었다[12].

CONSORT는 전 세계 수많은 학술지와 세계의학편집인협회(Word Association of Medical Editors), 국제의학학술지편집인위원회(International Committee of Medical Journal Editors), 미국 과학학술지편집인협의회(Council of Scientific Editors) 등 저명한 편집 관련 단체들의 지지를 받고 있다. 학술지 내 CONSORT 도입은 무작위 배정 임상시험 보고의 질 향상과 관련이 있는 것으로 나타났다. 일부 근거에 따르면, 학술지가 CONSORT를 지지하는 경우 보고의 질이 향상되며, 시간이 지남에 따라 보고가 더욱 개선되는 것으로 나타났다[2,13-15]. 16,604건의 임상시험에 대한 50건의 평가를 종합한 코크란 리뷰(Cochrane review)에서 학술지의 CONSORT 지지 여부와 학술지가 발표한 임상시험 보고 사이의 연관성을 평가했는데, 27개의 CONSORT 점검표 항목 중 25개 항목이 CONSORT를 지지하는 학술지에 발표된 임상시험에서 지지하지 않는 학술지보다 더 완전하게 보고되었다[2,14]. 그러나 인과 관계를 증명할 수는 없다. 최소한 CONSORT는 많은 최종 사용자(예: 저자, 학술지 편집인, 동료 심사자)에게 무작위 배정 임상시험에서 신중하고 철저한 보고가 얼마나 중요한지에 대한 인식을 일깨워주었다.

SPIRIT와 CONSORT는 각각 완료된 무작위 임상시험의 프로토콜과 1차 보고서에 포함되어야 하는 필수 항목의 점검표와, 임상시험 참여자의 흐름을 문서화한 도표로 구성된, 근거 기반 가이드라인이다. 이 지침은 임상시험 계획서와 보고서가 명확하고 투명하게 작성될 수 있도록, 임상시험 보고가 포함해야 하는 최소한의 정보에 대한 지침을 작성자에게 제공한다. 또한 각 점검표 항목의 의미와 근거, 좋은 보고의 예, 그리고 가능하다면 관련된 경험적 근거까지 제공하는 설명서 및 상세 문서를 함께 발행한다.

2020년 1월, 영국 옥스퍼드에서 SPIRIT과 CONSORT의 임원진이 모였다. 이들은 SPIRIT와 CONSORT statements는 개념적으로 연결되어 있고, 내용이 중복되며, 전파 및 이행 전략도 유사하기 때문에 두 그룹이 함께 일하는 것이 더 효과적이라고 판단하여 하나의 그룹을 구성하였다. CONSORT 2025 statement는 *The BMJ*, *JAMA*, *The Lancet*, *Nature Medicine*, *PLoS Medicine*에 동시에 게재될 예정이다.

SPIRIT 및 CONSORT statement 개정 결정

SPIRIT와 CONSORT는 생동하는(living) 가이드라인으로, 새로운 근거, 방법론적 발전, 사용자의 피드백을 반영하여 주기적으로 개정하는 것이 매우 중요하다. 그렇지 않으면, 시간이 지

남에 따라 그 가치와 유용성이 감소할 수 있다[16]. SPIRIT 2013과 CONSORT 2010 statement를 함께 개정하는 것은 두 점검표의 일관성을 더욱 높이고, 임상시험의 설계, 수행, 분석, 결과 등 임상시험 계획서부터 최종 출판에 이르기까지의 과정에서 사용자에게 일관된 지침을 제공할 수 있는 기회이기도 하다. 보고 과정을 조화롭게 만드는 것은 사용성과 준수도를 향상하고, 더 완전한 보고로 이어질 것이다[17]. 여기에서는 개정된 CONSORT 2025 statement를 소개하며, 개정된 SPIRIT 2025 statement는 별도로 발표되었다[18].

CONSORT 2025의 개발

CONSORT statement 개정에 사용된 방법은 보건 연구 지침 개발자를 위한 EQUATOR 네트워크 지침을 따랐으며[19], 이는 다른 곳에서 자세히 설명하였다[20,21]. 간단히 설명하자면, 먼저 문헌에 대한 범위 검토를 수행하여 수정 및 추가를 제안하거나 CONSORT 2010의 강점과 과제를 성찰하는 기존 출판 의견을 확인했으며, 그 결과는 별도로 발표되었다[22]. 또한 CONSORT 및 무작위 배정 임상시험의 편향 위험과 관련된 경험적, 이론적 증거를 위한 프로젝트별 데이터베이스(SCEBdb)를 개발하였다[23]. 범위 검토에서 확인된 근거는, 점검표 항목이 모든 임상시험(유해성[24], 결과[25]) 또는 상당수의 임상시험[26](비약물적 치료[27])에 적용되는 특정 주요 기존 CONSORT 확장(extension)의 주요 저자가 제공한 권고사항 및 근거와 결합되었으며, 기타 관련 보고 가이드라인(중재 설명 및 복제를 위한 템플릿[template for intervention description and replication, TIDieR][28]) 및 기타 출처(예: 개인 커뮤니케이션)의 권고와도 결합되었다.

기존 CONSORT 2010 점검표를 출발점으로 삼고, 범위 검토 및 권고사항에서 수집한 근거를 사용하여 점검표에 대한 잠재적 수정 또는 추가 목록을 작성하였다. 이 잠재적 변경사항 목록을 최종 사용자에게 제시하고, 1라운드에 317명, 2라운드에 303명, 3라운드에 290명이 응답한 대규모 국제 온라인 델파이 설문조사에서 피드백을 받았다. 델파이 참가자는 기존의 SPIRIT 및 CONSORT 협력체와 전문 연구 네트워크 및 학회를 통해 선정하였다. 또한 SPIRIT-CONSORT 개정 프로젝트 웹사이트에 관심 표명 양식을 마련하여 참가자를 모집하였다. 최종 사용자의 역할은 다양했는데, 통계학자/방법론자/역학자(n=198), 체계적 문헌고찰자/가이드라인 개발자(n=73), 임상시험 조사자(n=73), 임상의(n=58), 학술지 편집인(n=47), 환자 대표(n=17)가 가장 많았다(복수 소속 가능). 세 차례에 걸친 델파이 설문조사에서, 참가자들에게 개정된 CONSORT 점검표에 포함된 각 항목에 어느 정도 동의하는지 5점 리커트 척도로 평가하도록 요청하였다. 각 항목에 대한 의견을 제시하고 새로운 점검표 항목을 추가로 제안할 수 있는 무료 텍스트 상자도 제공하였다.

델파이 설문조사 결과는 2023년 3월 1일과 2일 이틀간 Zoom (Zoom Video Communications Inc.)을 통해 진행된 온라인 전문가 합의 회의에서 발표되고 논의되었는데, 여기에는 델파이 설문조사에 포함된 다양한 이해관계자 그룹을 대표하는 초청 국제 참가자 30명이 참석하였다. 이 회의에서는 새로 추가되거나 수정된 CONSORT 점검표 항목에 대해 논의하고 합의를 모색하였다. 의견 차이가 있는 항목에 대해서는 Zoom을 통한 익명 투표로 지지 수준을 파악하였다. 이러한 투표는 단지 권고적 성격으로, 공식적인 합의 임계값(threshold)은 지정되지 않았다.

전문가 합의 회의 이후, 실무진은 2023년 4월 25일과 26일 이틀간 옥스퍼드에서 대면 서면 회의를 개최하여 새롭게 추가되거나 수정된 CONSORT 점검표 각 항목의 형식과 문구를 검토하고 합의하였다. 그런 다음 점검표 초안을 컨센서스 회의 참가자들이 회람하고, 각 항목이 그룹 합의를 반영하는지 또는 추가 설명이 필요한지를 확인하였다. 이 피드백을 바탕으로 임원진에서 CONSORT 항목을 추가로 수정하였다. 최종 수정된 항목은 임상시험 보고서에 포함해야 할 최소한의 내용을 다루고 있으나, 이는 향후 저자들이 중요하다고 생각하거나 복제를 용이하게 하는 추가 정보를 포함하지 못하게 하는 것은 아니다. 집행 그룹의 구성원과 30명의 초청된 합의 회의 참가자는 본 원고의 저자이며, 이들의 이름은 원고 말미에 기재되어 있다.

CONSORT 2025의 주요 변경사항

CONSORT 2025 점검표에 여러 가지 실질적인 변경사항을 적용하였다(박스 1 참조). 7개의 새로운 점검표 항목을 추가하고, 3개 항목을 수정했으며, 1개 항목을 삭제했다. 아울러 주요 CONSORT 확장(유해성[24], 결과[25], 비약물학적 치료[27]) 및 기타 관련 보고 가이드라인(TIDieR [28])의 여러 항목을 통합하였다. 임상시험 등록(항목 2), 임상시험 프로토콜 및 통계 분석 계획에 접근할 수 있는 위치(항목 3), 비식별화된 참가자 수준의 데이터 공유(항목 4), 자금 및 이해관계(항목 5) 등 개념적으로 연결된 항목을 포함하는 오픈 사이언스에 대한 새로운 섹션을 추가하여 CONSORT 점검표도 재구성하였다. 또한 CONSORT와 SPIRIT 점검표 항목 간의 문구를 통일하고, 일부 항목의 문구를 명확하고 간결하게 다듬었다. CONSORT 2010 점검표와 CONSORT 2025 점검표의 변경사항을 자세히 비교하려면 Supplement 1을 참조하면 된다. 더불어 CONSORT 설명 및 상세화 문서를 개정하여[29], 각 CONSORT 2025 점검표 항목에 대한 근거와 과학적 배경을 설명하고 우수 보고 사례를 게시하였다.

박스 1. CONSORT 2025의 주요 변경사항 요약

새로운 점검표 항목 추가

- 항목 4: 비식별화된 개별 참여자 데이터, 통계 코드 및 기타 자료에 접근할 수 있는 위치와 방법을 포함한 데이터 공유에 관한 항목 추가
- 항목 5b: 원고 작성자의 재정적 및 기타 이해상충에 관한 항목 추가
- 항목 8: 임상시험의 설계, 수행 및/또는 보고에 환자 및/또는 대중이 어떻게 참여했는지에 대한 항목 추가
- 항목 12b: 해당되는 경우, 시험기관 및 중재를 제공하는 개인의 자격 기준에 대한 항목 추가
- 항목 15: 위해 및 기타 의도하지 않은 영향 평가방법에 대한 항목 추가
- 항목 21: 각 분석에 포함되는 대상(예: 모든 무작위 참가자) 및 그룹(항목 21b), 분석에서 누락된 데이터를 처리하는 방법(항목 21c)을 정의하는 항목 추가
- 항목 24: 중재와 비교군이 실제로 어떻게 시행되었는지에 대한 내용(항목 24a), 임상시험 기간 동안 받은 동반 치료의 세부 사항(항목 24b) 등 중재 제공에 관한 항목 추가

전면 수정된 점검표 항목

- 항목 3: 임상시험 계획서 외에 통계 분석계획에 접근할 수 있는 위치를 포함하도록 항목 수정
- 항목 10: 임상시험 시작 후 미리 지정되지 않은 결과 또는 분석을 포함한 임상시험의 중요한 변경사항을 보고에 포함하도록 항목 수정
- 항목 26: 1차 및 2차 각 결과별로 분석에 포함된 참가자 수와, 각 시점에 사용 가능한 데이터를 가진 각 치료군별 참가자 수를 명시하도록 항목 수정

점검표 항목 삭제

- 임상시험 결과의 일반화 가능성에 대한 항목 삭제(임상시험 제한사항[30번 항목]에 통합)

주요 CONSORT 확장의 점검표 항목 통합

- 위해[24]를 평가 및 분석한 방법(항목 7, 15, 21a, 23a, 27), 결과[25]를 측정 및 분석한 방법(항목 14, 26), 중재[27,28]와 비교군이 실제로 어떻게 시행되고 누가 시행했는지(항목 24) 보고와 관련된 항목 추가

점검표 항목의 구조 및 구성

- 임상시험 등록(항목 2), 임상시험 프로토콜 및 통계 분석계획에 접근할 수 있는 위치(항목 3), 비식별화된 참가자 수준의 데이터 공유(항목 4), 자금 및 이해상충(항목 5) 등 개념적으로 연결된 항목을 포함한 오픈 사이언스에 대한 섹션을 신설하여 점검표를 재구성
- 일부 CONSORT 점검표 항목의 문구를 SPIRIT 점검표 항목의 문구와 통일하고, 그 반대도 적용
- 일부 항목의 문구를 명확하고 간결하게 수정

또한 CONSORT 2025의 원활한 이행을 돕기 위해, 각 항목의 핵심 요소를 도출하여 글머리 기호 형식으로 정리한 확장 버전 CONSORT 2025 점검표를 개발하였다. 이는 COBWEB (CONSORT 기반 웹 도구) [30] 및 COBPeer (CONSORT 기반 동료 검토 도구) [31] 연구에서 제안하고 체계적 문헌고찰 보고를 위한 2020 PRISMA 지침[32]에서 사용한 모델과 유사하다. 확장된 점검표는 CONSORT 2025 설명 및 상세화 문서에 제시된 요소를 요약한 버전으로[29], 예시와 참조는 제외되었다(Supplement 2).

CONSORT 2025의 범위

CONSORT 2025 statement는 30개 항목의 점검표로 구성되어 있으며, 무작위 배정 임상시험 보고서에 포함되어야 할 최소한의 항목(표 1)과 임상시험 참여자의 흐름을 문서화하는 도표를 제공한다(그림 1). CONSORT 2025 statement는 CONSORT 2025 설명 및 상세 문서와 함께 사용할 것을 강력히 권장한다[29]. CONSORT 2025 statement는 CONSORT 2010 statement를 대체

표 1. 무작위 임상시험을 보고할 때 포함해야 할 정보에 대한 CONSORT 2025 체크리스트

섹션/주제	No.	CONSORT 2025 체크리스트 항목 설명
제목 및 초록		
제목 및 구조화된 초록	1a	무작위 임상시험으로서의 식별
	1b	임상시험 설계, 방법, 결과 및 결론의 구조화된 요약
오픈 사이언스		
임상시험 등록	2	임상시험 등록부 이름, 식별번호(URL 포함) 및 등록날짜
프로토콜 및 통계 분석계획	3	시험 프로토콜 및 통계 분석계획에 접근할 수 있는 위치
데이터 공유	4	비식별화된 개별 참가자 데이터(데이터 사전 포함), 통계 코드 및 기타 자료에 접근할 수 있는 위치 및 방법
자금 및 이해관계	5a	자금 및 기타 지원 출처(예: 의약품 공급) 및 시험의 설계, 수행, 분석 및 보고에서 자금 제공자의 역할
	5b	원고 저자의 재정 및 기타 이해관계

(Continued on the next page)

표 1. Continued

섹션/주제	No.	CONSORT 2025 체크리스트 항목 설명
서론		
배경 및 근거	6	과학적 배경 및 근거
목적	7	이익 및 위해와 관련된 구체적인 목표
방법		
환자 및 대중 참여	8	시험의 설계, 수행 및 보고에 대한 환자 또는 대중 참여 세부 사항
시험 설계	9	시험 유형을 포함한 시험 설계 설명(예: 평행군, 교차), 할당 비율 및 프레임워크(예: 우월성, 동등성, 비열등성, 탐색적)
시험 프로토콜 변경	10	시험 시작 후 미리 지정되지 않은 결과 또는 분석을 포함한, 시험에 대한 중요한 변경사항 및 이유
시험 설정	11	설정(예: 지역사회, 병원) 및 위치(예: 국가, 시험기관)
자격 기준	12a	참가자 자격 기준
	12b	해당되는 경우, 기관 및 중재를 제공하는 개인(예: 외과의, 물리치료사)에 대한 자격 기준
중재 및 비교군	13	복제를 허용하기에 충분한 세부 정보가 포함된 중재 및 비교군. 해당되는 경우, 중재 및 비교자를 설명하는 추가 자료(예: 중재 매뉴얼)에 접근할 수 있는 곳
결과	14	특정 측정 변수(예: 수축기 혈압), 분석 지표(예: 기준치, 최종값, 이벤트까지의 시간 변화), 집계방법(예: 중앙값, 비율) 및 각 결과에 대한 시점
위해	15	위해를 정의하고 평가한 방법(예: 체계적, 비체계적)
표본 크기	16a	표본 크기 계산을 뒷받침하는 모든 가정을 포함하여 표본 크기를 결정한 방법
	16b	중간 분석 및 중단 지침에 대한 설명
무작위 배정:		
시퀀스 생성	17a	무작위 할당 시퀀스를 생성한 사람 및 사용된 방법
	17b	무작위 할당 유형 및 제한 사항(예: 층화, 차단 및 블록 크기)
할당 은폐 메커니즘	18	무작위 할당 시퀀스를 구현하는 데 사용된 메커니즘(예: 중앙 컴퓨터/전화, 순차적으로 번호가 매겨진 불투명하고 밀폐된 용기), 중재가 배정될 때까지 순서를 은폐하는 모든 단계를 설명
구현	19	등록한 직원과 참여자를 중재에 배정하는 직원이 무작위 배정 순서에 접근할 수 있는지 여부
눈가림	20a	중재 배정 후의 눈가림 대상(예: 참가자, 의료 제공자, 결과 평가자, 데이터 분석가)
	20b	눈가림이 이루어진 경우, 눈가림이 어떻게 시행되었는지와 중재 간 유사성에 대한 설명
통계방법	21a	위해를 포함한 일차 및 이차 결과에 대한 그룹을 비교하는 데 사용한 통계적 방법
	21b	각 분석에 포함되는 사람(예: 모든 무작위 참가자) 및 그룹에 대한 정의
	21c	분석에서 누락된 데이터를 처리한 방법
	21d	추가 분석을 위한 방법(예: 하위 그룹 및 민감도 분석), 사전 지정과 사후 분석을 구분
결과		
흐름도를 포함한 참가자 흐름	22a	각 그룹에 대해 무작위로 배정되고, 의도된 중재를 받고, 일차 결과 분석에 포함된 참가자 수
	22b	각 그룹에 대해 무작위 배정 후의 손실 및 제외와 그 이유
모집	23a	혜택 및 피해 결과에 대한 모집 및 추적 기간을 정의하는 날짜
	23b	관련이 있는 경우 시험이 종료되거나 중단된 이유
중재 및 비교 대상 제공	24a	실제로 시행된 중재 및 비교 대상(예: 적절한 경우, 누가 중재/비교약을 제공했는지, 참가자가 어떻게 준수했는지, 의도대로 제공되었는지 여부[충실도])
	24b	각 그룹이 시험기간에 받은 병용 치료
기준 데이터	25	각 그룹의 기준 인구통계 및 임상 특성을 보여주는 표
분석된 숫자, 결과 및 추정	26	각 1차 및 2차 결과, 그룹별로: - 분석에 포함된 참가자 수 - 결과 시점에 사용 가능한 데이터를 가진 참가자 수 - 각 그룹에 대한 결과 및 추정 효과 크기와 그 사전값
위해	27	각 그룹에서 발생한 모든 위해 또는 의도치 않은 사건
보조 분석	28	사전 지정 분석과 사후 분석을 구별하여 수행된 기타 분석(하위 그룹 및 민감도 분석 포함)
고찰		
해석	29	결과에 부합하고, 이점과 위해의 균형을 맞추며, 기타 관련 근거를 고려한 해석
한계	30	잠재적 편향, 부정확성, 일반화 가능성 및 관련 있는 경우, 분석의 다중성 원인을 다루는 시험 한계

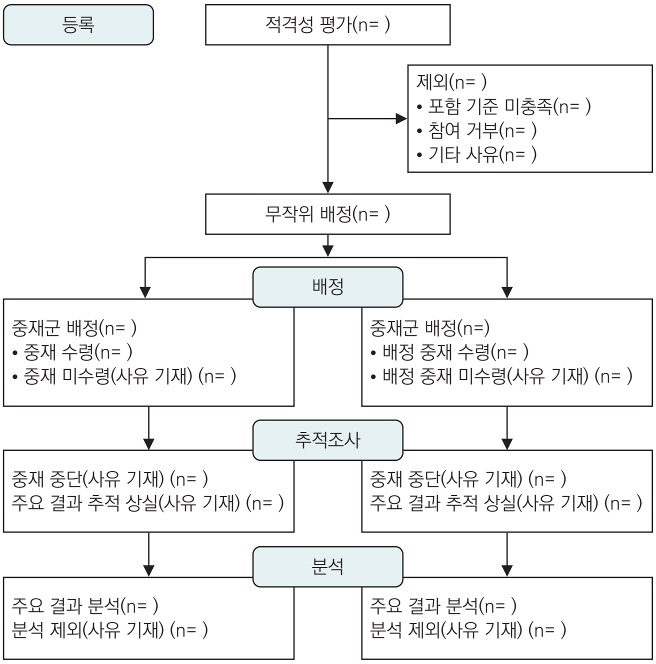

등록
적격성 평가(n=)
제외(n=)
• 포함 기준 미충족(n=)
• 참여 거부(n=)
• 기타 사유(n=)
무작위 배정(n=)
배정
중재군 배정(n=)
• 중재 수령(n=)
• 중재 미수령(사유 기재) (n=)
중재군 배정(n=)
• 배정 중재 수령(n=)
• 배정 중재 미수령(사유 기재) (n=)
추적조사
중재 중단(사유 기재) (n=)
주요 결과 추적 상실(사유 기재) (n=)
중재 중단(사유 기재) (n=)
주요 결과 추적 상실(사유 기재) (n=)
분석
주요 결과 분석(n=)
분석 제외(사유 기재) (n=)
주요 결과 분석(n=)
분석 제외(사유 기재) (n=)

그림 1. CONSORT 2025 흐름도. 두 그룹 무작위 배정 임상시험의 단계별 진행 상황(즉 등록, 중재 할당, 추적 관찰 및 데이터 분석)을 나타내는 흐름도. CONSORT, Consolidated Standards of Reporting Trials.

하므로, CONSORT 2010 statement는 더 이상 사용되어서는 안 된다. 학술지 편집인과 출판사는 저자 지침을 개정하여 CONSORT 2025를 참조하도록 해야 한다. CONSORT 2025는 모든 무작위 배정 임상시험 보고에 대한 지침을 제공하지만, 가장 일반적인 유형인 두 집단 병렬 설계(two-group parallel design)에 중점을 둔다(그림 1).

CONSORT의 확장판은 다양한 유형의 임상시험 설계, 데이터 및 중재유형의 보고와 관련된 방법론적 문제를 해결하기 위해 개발되었다. 임상시험 설계 확장에는 적응 설계[33], 군집 시험[34], 교차 시험[35], 초기 단계 시험[36], 요인 시험[37], 비열등성 및 동등성 시험[38], 실용적 시험[39], 다군 시험[40], n-of-1 시험[41], 시범 및 타당성 시험[42], 사람 내 무작위 배정 시험[43]에 대한 권장사항이 포함된다. 그 외 비약물학적 치료[27], 결과[25], 환자 보고 결과[44], 대리 결과[45], 사회적 및 심리적 개입[46], 피해[24], 초록[47], 건강 형평성[48] 등이 포함된다. 우리는 이러한 확장판의 책임자들과 협력하여, 개정된 CONSORT 2025 statement에 부합하는 프로세스를 구현할 것이다. 독자들은 당분간 기존 버전의 관련 CONSORT 확장판을 사용할 것을 권장한다.

시사점 및 한계

CONSORT 2025 statement의 목적은 명확하고 완전하며 투명한 방식으로 임상시험을 보고하기 위해 저자가 포함해야 하는 내용에 대한 최소한의 권장사항을 제공하는 것이다[9,10]. 독자, 동료 심사자, 임상의, 가이드라인 작성자, 환자 및 일반인, 그리고 편집인 또한 CONSORT 2025를 활용하여 무작위 배정 임상시험의 보고를 평가하는 데 도움을 받을 수 있다. 또한 원고 제출과정에서, 각 점검표 항목이 원고의 어디에 보고되어 있는지를 명시한 완성된 CONSORT 2025 점검표를 제출하고, 이를 보충자료의 일부로 업로드할 것을 강력히 권장한다[49]. 모호함이나 누락 없이 수행한 작업과 발견한 내용을 명시적으로 설명하는 것이 모든 독자의 이익에 가장 부합한다[9].

CONSORT 2025와 SPIRIT 2025에는 임상시험 설계, 수행 또는 분석에 대한 권장사항을 포함하고 있지 않지만, 그럼에도 불구하고 여기에 포함된 권장사항은 고려해야 할 주요 이슈를 강조함으로써 연구자들이 임상시험을 설계, 수행 및 분석하는 데 도움을 줄 수 있다. 그리고 SPIRIT 및 CONSORT statement를 함께 개정하는 것은 두 점검표의 보고 내용을 일치시키고 임상시험 계획서부터 최종 출판까지 임상시험 설계, 수행 및 분석 보고에 대한 일관된 지침을 사용자에게 제공할 수 있는 기회이기도 했다[17]. 따라서 임상시험 계획서의 명확하고 투명한 보고는 결과적으로 적절하게 설계되고 잘 수행된 임상시험을 촉진하는 데 도움이 될 것이다. 또한 임상시험 결과를 투명하게 보고하면, 연구 결함이 존재하는 경우 이를 밝혀내고 그 유병률과 심각성을 더 잘 추정할 수 있다. 그러나 중요한 것은 CONSORT 2025가 품질 평가도구로 사용되지는 않는다는 점이다. CONSORT 2025의 내용은 오히려 무작위 배정 임상시험의 내부 및 외부 유효성과 관련된 보고 항목에 초점을 맞추고 있다.

CONSORT 2025에서는 무작위 임상시험 보고를 위한 엄격한 구조를 제안하지 않는다. 대신, 논문의 형식은 학술지의 개별 스타일과 “저자를 위한 지침”을 준수해야 한다. 저자는 논문의 어딘가에 점검표 항목을 충분히 상세하고 명확하게 언급해야 한다[9]. 또한 임상시험 방법과 결과를 일부 인쇄 학술지 논문의 일반적인 분량 한도보다 더 자세히 보고할 수 있도록, 추가적인 온라인 보충자료의 사용을 권장한다. 전체 데이터 및 코드 공유는 또 다른, 더 높은 수준의 투명성을 제공하며, 무작위 임상시험에서 이러한 일이 발생했는지 또는 발생할 계획인지(예: 일정 시간 후)에 대한 자세한 정보를 제공할 것을 권장한다.

CONSORT는 실제 임상시험 설계와 수행, 분석을 반영하는 명확하고 투명한 보고를 촉구한다. 고품질의 보고는 재현성과 관련된 문제를 고려할 때 중요한 단계이다[50]. 임상시험 저자

는 수행된 사항을 자세히 설명하고, 수행되지 않았거나 수정된 사항이 있는 경우, 이를 인정하여 임상시험 계획서, 통계 분석 계획서 및 임상시험 등록부에 보고된 정보와 일치하도록 하는 것이 좋다. 연구자, 연구 수련생, 학술지 편집인, 동료 심사자를 대상으로 하는 추가 리소스 및 교육 자료를 포함하여 CONSORT 및 SPIRIT statement에 대한 자세한 정보를 제공하기 위해 SPIRIT-CONSORT 공동 웹사이트(https://www.consort-spirit.org/)가 개설되었다. 이 웹사이트에는 환자와 대중을 대상으로 한 자료도 포함되어 있어, 무작위 배정 임상시험의 명확하고 투명한 보고의 중요성과 근거 기반 의료 제공에서의 중요성을 설명한다.

CONSORT 2025는 새로운 증거와 새로운 관점을 반영하기 위해 주기적으로 개정되는, 생동하는(living) 가이드라인이다. 이러한 접근방식은 저자, 환자 및 일반인, 학술지 편집인, 동료 심사자 등 최종 사용자에게 지침의 관련성을 유지하는 데 중요하다.

Additional information

[1]Oxford Clinical Trials Research Unit, Centre for Statistics in Medicine, University of Oxford, Oxford, UK
[2]Department of Medicine, Women's College Research Institute, University of Toronto, Toronto, ON, Canada
[3]United Kingdom EQUATOR Centre, Centre for Statistics in Medicine, University of Oxford, Oxford, UK
[4]Department of Clinical Research, Centre for Evidence-Based Medicine Odense and Cochrane Denmark, University of Southern Denmark, Odense, Denmark
[5]Open Patient data Explorative Network, Odense University Hospital, Odense, Denmark
[6]Centre for Journalology, Clinical Epidemiology Programme, Ottawa Hospital Research Institute, Ottawa, ON, Canada
[7]Department of Obstetrics and Gynecology, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[8]Jawaharlal Institute of Postgraduate Medical Education and Research, Puducherry, India
[9]Office of Science Dissemination, Centers for Disease Control and Prevention, Atlanta, GA, USA
[10]Department of Biostatistics and Epidemiology, School of Public Health, Center for Pharmacoepidemiology and Treatment Science, Rutgers University, New Brunswick, NJ, USA
[11]JAMA Network Open, Chicago, IL, USA
[12]Centre for Health Research and Development, Society for Applied Studies, New Delhi, India
[13]Child Health Evaluation Services, The Hospital for Sick Children Research Institute, Toronto, ON, Canada
[14]Department of Psychiatry, University of Toronto, Toronto, ON, Canada
[15]Aberdeen Centre for Evaluation, University of Aberdeen, Aberdeen, UK
[16]Project PINK BLUE-Health & Psychological Trust Centre, Abuja, Nigeria
[17]Department of Sociology and Gerontology, Miami University, Oxford, OH, USA
[18]Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, UK
[19]Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK
[20]Ottawa Hospital Research Institute, Ottawa, ON, Canada
[21]Department of Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA
[22]Department of Epidemiology and Population Health, Stanford University, Palo Alto, CA, USA
[23]Institute for Evidence-Based Healthcare, Faculty of Health Sciences and Medicine, Bond University, Robina, QLD, Australia
[24]Departments of Medicine, of Epidemiology and Population Health, of Biomedical Data Science, and of Statistics, and Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, CA, USA
[25]MRC Clinical Trials Unit at University College London, London, UK
[26]University College London, UCL Great Ormond Street Institute of Child Health, London, UK
[27]NIHR Exeter Biomedical Research Centre, Faculty of Health and Life Sciences, University of Exeter, Exeter, UK
[28]Edinburgh Clinical Trials Unit, Usher Institute-University of Edinburgh, Edinburgh, UK
[29]The BMJ, BMA House, London, UK
[30]Harvard Medical School, Boston, MA, USA
[31]Université Paris Cité, Inserm, INRAE, Centre de Recherche Epidémiologie et Statistiques, Université Paris Cité, Paris, France
[32]Clinical Trials Ontario, MaRS Centre, Toronto, ON, Canada[33]Duke Clinical Research Institute, Duke University Medical Center, Durham, NC, USA
[34]Department of Emergency Medicine, University of California, Los Angeles, CA, USA
[35]South African Medical Research Council, Cape Town, South Africa
[36]Warwick Applied Health, Warwick Medical School, University of Warwick, Coventry, UK
[37]MRC/CSO Social and Public Health Sciences Unit & Robertson Centre for Biostatistics, Institute of Health and Wellbeing, University of Glasgow, Glasgow, UK
[38]Department of Health Research Methods Evidence and Impact, McMaster University, Hamilton, ON, Canada
[39]St. Joseph's Healthcare Hamilton, Hamilton, ON, Canada
[40]York Trials Unit, Department of Health Sciences, University of York, York, UK
[41]Faculty of Medicine and Dentistry, University of Alberta, Edmonton, AB, Canada
[42]Université Paris Cité and Université Sorbonne Paris Nord, Inserm, INRAE, Centre for Research in Epidemiology and Statistics (CRESS), Paris, France
[43]Centre d'Epidémiologie Clinique, Hôpital Hôtel Dieu, AP-HP, Paris, France

ORCID

Sally Hopewell: https://orcid.org/0000-0002-6881-6984
An-Wen Chan: https://orcid.org/0000-0002-4498-3382
Gary S. Collins: https://orcid.org/0000-0002-2772-2316
Asbjørn Hróbjartsson: https://orcid.org/0000-0002-2451-5012
David Moher: https://orcid.org/0000-0003-2434-4206
Kenneth F. Schulz: https://orcid.org/0000-0003-2596-7616
Ruth Tunn: https://orcid.org/0000-0003-1187-1808
Rakesh Aggarwal: https://orcid.org/0000-0001-9689-494X
Michael Berkwits: https://orcid.org/0009-0005-1316-5861
Jesse A. Berlin: https://orcid.org/0000-0002-9810-745X
Nita Bhandari: https://orcid.org/0000-0003-0349-087X
Nancy J. Butcher: https://orcid.org/0000-0002-5152-0108

Marion K. Campbell: https://orcid.org/0000-0001-5386-4097

Runcie C. W. Chidebe: https://orcid.org/0000-0002-6025-776X

Diana Elbourne: https://orcid.org/0000-0003-3044-4545

Andrew Farmer: https://orcid.org/0000-0002-6170-4402

Dean A. Fergusson: https://orcid.org/0000-0002-3389-2485

Robert M. Golub: https://orcid.org/0009-0000-3270-0632

Tammy C. Hoffmann: https://orcid.org/0000-0001-5210-8548

John P. A. Ioannidis: https://orcid.org/0000-0003-3118-6859

Rachel L. Knowles: https://orcid.org/0000-0002-5490-7682

Sarah E. Lamb: https://orcid.org/0000-0003-4349-7195

Martin Offringa: https://orcid.org/0000-0002-4402-5299

Dawn P. Richards: https://orcid.org/0000-0003-1151-0826

Frank W. Rockhold: https://orcid.org/0000-0003-3732-4765

Nandi L. Siegried: https://orcid.org/0000-0002-4081-1698

Sophie Staniszewska: https://orcid.org/0000-0002-7723-9074

Rod S. Taylor: https://orcid.org/0000-0002-3043-6011

Lehana Thabane: https://orcid.org/0000-0003-0355-9734

David Torgerson: https://orcid.org/0000-0002-1667-4275

Sunita Vohra: https://orcid.org/0000-0002-6210-7933

Ian R. White: https://orcid.org/0000-0002-6718-7661

Isabelle Boutron: https://orcid.org/0000-0002-5263-6241

Ethical approval

Ethical approval was granted by the Central University Research Ethics Committee, University of Oxford (R76421/RE001). All Delphi participants provided informed consent to participate.

Authors' contribution

Conceptualization: SH, AWC, GSC, AH, DM, KFS, RT, RA, MB, NB, NJB, MKC, RCWC, DE, DAF, RMG, SNG, TCH, JPAI, BCK, RLK, SEL, SL, EL, MO, PR, DPR, FWR, DLS, NLS, SS, RST, LT, DT, SV, IRW, IB. Funding acquisition: SH, GSC. Investigation: SH, AWC, GSC, AH, DM, KFS, RT, RA, MB, NB, NJB, MKC, RCWC, DE, DAF, RMG, SNG, TCH, JPAI, BCK, RLK, SEL SL, EL, MO, PR, DPR, FWR, DLS, NLS, SS, RST, LT, DT, SV, IRW, IB. Methodology: SH, AWC, GSC, AH, DM, KFS, RT, RA, MB, NB, NJB, MKC, RCWC, DE, DAF, RMG, SNG, TCH, JPAI, BCK, RLK, SEL, SL, EL, MO, PR, DPR, FWR, DLS, NLS, SS, RST, LT, DT, SV, IRW, IB. Project administration: SH, RT. Supervision: SH, AWC, GSC, DM, KFS, IB. Writing–original draft: SH, AWC, GSC, AH, DM, KFS, RT, IB. Writing–review & editing: SH, AWC, GSC, AH, DM, KFS, RT, RA, MB, JAB, NB, NJB, MKC, RCWC, DE, AF, DAF, RMG, SNG, TCH, JPAI, BCK, RLK, SEL, SL, EL, MO, PR, DPR, FWR, DLS, NLS, SS, RST, LT, DT, SV, IRW, IB.

Conflict of interest

I have read the journal's policy and the authors of this manuscript have the following competing interests: support from MRC-NIHR for the submitted work. SH, IB, AWC, AH, KFS, and DM are members of the SPIRIT-CONSORT executive group. SH, IB, AWC, AH, KFS, GSC, DM, MKC, NJB, MO, RST, and SV are involved in the development, update, implementation, and dissemination of several reporting guidelines. GSC is the director of the UK EQUATOR Centre, a statistical editor for The BMJ and NIHR Senior Investigator, DM is the director of the Canadian EQUATOR Centre, and member of The BMJ's regional advisory board for North America, IB is deputy director and PR is director of the French EQUATOR Centre, TCH is director of the Australasian EQUATOR Centre, JPAI is director of the US EQUATOR Centre. RA is president of the World Association of Medical Editors. MKC is chair of the MRC-NIHR: Better Methods Better Research funding panel. RCWC is executive director of Project PINK-BLUE, which receives funding from Roche-Product. AF is director of the UK National Institute for Health and Care Research Health Technology Assessment Programme. DPR is a full time employee of Five02 Laboratories, which under contract to Clinical Trials Ontario provides services related to patient and public engagement; and is the volunteer vice president of the Canadian Arthritis Patient Alliance, which receives funding through independent grants from pharmaceutical companies. IRW was supported by the MRC Programmes MCUU00004/07 and MCUU00004/09. DLS is JAMA associate editor and receives editing stipends from JAMA and Annals of Emergency Medicine.

Funding

The 2025 update of SPIRIT and CONSORT was funded by the MRC-NIHR: Better Methods, Better Research (MR/W020483/1). The funder reviewed the design of the study but had no role in the collection, analysis, or interpretation of the data; in the writing of the report; or in the decision to submit the article for publication. The views expressed are those of the author(s) and not necessarily those of the NIHR, the MRC, or the Department of Health and Social Care.

Data availability

Not applicable.

Acknowledgments

SPIRIT-CONSORT executive group: Sally Hopewell, University of Oxford, UK; An-Wen Chan, University of Toronto, Canada; Asbjørn Hróbjartsson, University of Southern Denmark, Odense, Denmark; David Moher, University of Ottawa, Canada; Kenneth Schulz, University of North Carolina at Chapel Hill, USA; Isabelle Boutron, Université Paris Cité, France. Gary Collins and Ruth Tunn, both of University of Oxford, UK, were also involved in leading the SPIRIT and CONSORT 2025 update. SPIRIT-CONSORT 2025 consensus meeting participants: Rakesh Aggarwal, Jawaharlal Institute of Postgraduate Medical Education and Research, India; Michael Berkwits, Office of the Centers for Disease Control and Prevention, USA (formally JAMA and the JAMA Network at time of consensus meeting); Jesse A Berlin, Rutgers University/JAMA Network Open USA; Nita Bhandari, Society for Applied Studies, India; Nancy J Butcher, The Hospital for Sick Children, Canada; Marion K Campbell, University of Aberdeen, UK; Runcie W Chidebe, Project PINK BLUE, Nigeria/Miami University, Ohio, USA; Diana Elbourne, London School of Hygiene and Tropical Medicine, UK; Andrew J Farmer, University of Oxford, UK; Dean A Fergusson, Ottawa Hospital Research Institute, Canada; Robert M Golub, Northwestern University, USA; Steven N Goodman, Stanford University, USA; Tammy C Hoffmann, Bond University, Australia; John PA Ioannidis, Stanford University, USA; Brennan C Kahan, University College London, UK; Rachel L Knowles, University College London, UK; Sarah E Lamb, University of Exeter, UK; Steff Lewis, University of Edinburgh, UK; Elizabeth Loder, The BMJ, UK; Martin Offringa, Hospital for Sick Children Research Institute, Canada; Dawn P Richards, Clinical Trials Ontario, Canada; Frank W Rockhold, Duke University, USA; David L Schriger, University of California, USA; Nandi L Siegfried, South African Medical Research Council, South Africa; Sophie Staniszewska, University of Warwick, UK; Rod S Taylor, University of Glasgow, UK; Lehana Thabane, McMaster University/St. Joseph's Healthcare, Canada; David Torgerson, University of York, UK; Sunita Vohra, University of Alberta, Canada; and Ian R White, University College London, UK. We dedicate CONSORT 2025 to the late Doug Altman who was instrumental in the development of the SPIRIT and CONSORT statements and whose files and correspondence contributed to this update following his death. We gratefully acknowledge the contributions of all those who participated in the SPIRIT-CONSORT 2025 update Delphi survey. We also acknowledge Camilla Hansen Nejstgaard for conducting the scoping review of the literature to identify suggested changes to CONSORT 2010 and for comparing CONSORT 2025 with RoB2, Lasse Østengaard for developing the SCEB (SPIRIT-CONSORT Evidence Bibliographic) database of empirical evidence to support the development of CONSORT, Jen de Bayer and Patricia Logullo for their involvement during the consensus meeting, and Lina El Ghosn for initial drafting of the expanded CONSORT checklist.

Patient and public involvement: The CONSORT 2025 checklist items and the explanations here were developed using input from an international Delphi survey and consensus meeting. The Delphi survey was advertised via established patient and public involvement (PPI) networks, and 17 respondents self-identified as a "patient or public representative" and completed the Delphi survey. In addition, three of the participants in the expert consensus meeting were patient or public representatives who were leaders in advancing PPI. Dissemination to participants and related patient and public communities: CONSORT 2025 will be disseminated via a new website, consort-spirit.org, which will include materials designed for patients and the public.

Supplementary materials

Supplementary files are available from https://doi.org/10.7910/DVN/F1OE6L

Supplement 1. Comparison of CONSORT 2025 and CONSORT 2010 checklists.

Supplement 2. CONSORT 2025 expanded checklist of detailed information to include when reporting a randomised trial.

References

1. Altman DG. Better reporting of randomised controlled trials: the CONSORT statement. BMJ 1996;313:570-571. https://doi.org/10.1136/bmj.313.7057.570
2. Turner L, Shamseer L, Altman DG, Schulz KF, Moher D. Does use of the CONSORT Statement impact the completeness of reporting of randomised controlled trials published in medical journals?: a Cochrane review. Syst Rev 2012;1:60. https://doi.org/10.1186/2046-4053-1-60
3. Glasziou P, Altman DG, Bossuyt P, Boutron I, Clarke M, Julious S, Michie S, Moher D, Wager E. Reducing waste from incom-

plete or unusable reports of biomedical research. Lancet 2014; 383:267-276. https://doi.org/10.1016/S0140-6736(13)62228-X

4. Savovic J, Jones HE, Altman DG, Harris RJ, Juni P, Pildal J, Als-Nielsen B, Balk EM, Gluud C, Gluud LL, Ioannidis JP, Schulz KF, Beynon R, Welton NJ, Wood L, Moher D, Deeks JJ, Sterne JA. Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. Ann Intern Med 2012;157:429-438. https://doi.org/10.7326/0003-4819-157-6-201209180-00537
5. Goldacre B, Drysdale H, Powell-Smith A, Dale A, Milosevic I, Slade E, Hartley P, Marston C, Mahtani K, Heneghan C. The COMPare trials project [Internet]. Bennett Institute for Applied Data Science; 2016 [cited 2021 May 26]. Available from: https://www.compare-trials.org/
6. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, Pitkin R, Rennie D, Schulz KF, Simel D, Stroup DF. Improving the quality of reporting of randomized controlled trials: the CONSORT statement. JAMA 1996;276:637-639. https://doi.org/10.1001/jama.276.8.637
7. Moher D, Schulz KF, Altman D; CONSORT Group (Consolidated Standards of Reporting Trials). The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. JAMA 2001;285:1987-1991. https://doi.org/10.1001/jama.285.15.1987
8. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, Gotzsche PC, Lang T; CONSORT GROUP (Consolidated Standards of Reporting Trials). The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med 2001;134:663-694. https://doi.org/10.7326/0003-4819-134-8-200104170-00012
9. Schulz KF, Altman DG, Moher D; CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. Ann Intern Med 2010;152:726-732. https://doi.org/10.7326/0003-4819-152-11-201006010-00232
10. Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, Elbourne D, Egger M, Altman DG. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010;340:c869. https://doi.org/10.1136/bmj.c869
11. Chan AW, Tetzlaff JM, Altman DG, Laupacis A, Gotzsche PC, Krleza-Jeric K, Hrobjartsson A, Mann H, Dickersin K, Berlin JA, Dore CJ, Parulekar WR, Summerskill WS, Groves T, Schulz KF, Sox HC, Rockhold FW, Rennie D, Moher D. SPIRIT 2013 statement: defining standard protocol items for clinical trials. Ann Intern Med 2013;158:200-207. https://doi.org/10.7326/0003-4819-158-3-201302050-00583
12. Chan AW, Tetzlaff JM, Gotzsche PC, Altman DG, Mann H, Berlin JA, Dickersin K, Hrobjartsson A, Schulz KF, Parulekar WR, Krleza-Jeric K, Laupacis A, Moher D. SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. BMJ 2013;346:e7586. https://doi.org/10.1136/bmj.e7586
13. Plint AC, Moher D, Morrison A, Schulz K, Altman DG, Hill C, Gaboury I. Does the CONSORT checklist improve the quality of reports of randomised controlled trials?: a systematic review. Med J Aust 2006;185:263-267. https://doi.org/10.5694/j.1326-5377.2006.tb00557.x
14. Dechartres A, Trinquart L, Atal I, Moher D, Dickersin K, Boutron I, Perrodeau E, Altman DG, Ravaud P. Evolution of poor reporting and inadequate methods over time in 20920 randomised controlled trials included in Cochrane reviews: research on research study. BMJ 2017;357:j2490. https://doi.org/10.1136/bmj.j2490
15. Moher D, Jones A, Lepage L; CONSORT Group (Consolidated Standards for Reporting of Trials). Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation. JAMA 2001;285:1992-1995. https://doi.org/10.1001/jama.285.15.1992
16. Simera I, Altman DG. ACP Journal Club. Editorial: writing a research article that is “fit for purpose”: EQUATOR network and reporting guidelines. Ann Intern Med 2009;151:JC2-2-JC2-3. https://doi.org/10.7326/0003-4819-151-4-200908180-02002
17. Hopewell S, Boutron I, Chan AW, Collins GS, de Beyer JA, Hrobjartsson A, Nejstgaard CH, Ostengaard L, Schulz KF, Tunn R, Moher D. An update to SPIRIT and CONSORT reporting guidelines to enhance transparency in randomized trials. Nat Med 2022;28:1740-1743. https://doi.org/10.1038/s41591-022-01989-8
18. Chan AW, Boutron I, Hopewell S, Moher D, Schulz KF, Collins GS, Tunn R, Aggarwal R, Berkwits M, Berlin JA, Bhandari N, Butcher NJ, Campbell MK, Chidebe RC, Elbourne DR, Farmer AJ, Fergusson DA, Golub RM, Goodman SN, Hoffmann TC, Ioannidis JP, Kahan BC, Knowles RL, Lamb SE, Lewis S, Loder E, Offringa M, Ravaud P, Richards DP, Rockhold FW, Schriger DL, Siegfried NL, Staniszewska S, Taylor RS, Thabane

L, Torgerson DJ, Vohra S, White IR, Hrobjartsson A. SPIRIT 2025 statement: updated guideline for protocols of randomised trials. BMJ 2025;389:e081477. https://doi.org/10.1136/bmj-2024-081477

19. Moher D, Schulz KF, Simera I, Altman DG. Guidance for developers of health research reporting guidelines. PLoS Med 2010;7:e1000217. https://doi.org/10.1371/journal.pmed.1000217

20. Hopewell S, Chan AW, Boutron I, Hrobjartsson A, Collins G, Tunn R, de Beyer JA, Schulz KF, Moher D. Protocol for updating the SPIRIT 2013 (Standard Protocol Items: Recommendations for Interventional Trials) and CONSORT 2010 (CONsolidated Standards Of Reporting Trials) statements: version 1.0 [Internet]. OSF; 2022 [cited 2025 Apr 10]. Available from: https://doi.org/10.17605/OSF.IO/6HJYG

21. Tunn R, Boutron I, Chan AW, Collins GS, Hrobjartsson A, Moher D, Schulz KF, de Beyer JA, Hansen Nejstgaard C, Ostengaard L, Hopewell S. Methods used to develop the SPIRIT 2024 and CONSORT 2024 Statements. J Clin Epidemiol 2024;169:111309. https://doi.org/10.1016/j.jclinepi.2024.111309

22. Nejstgaard CH, Boutron I, Chan AW, Chow R, Hopewell S, Masalkhi M, Moher D, Schulz KF, Shlobin NA, Ostengaard L, Hrobjartsson A. A scoping review identifies multiple comments suggesting modifications to SPIRIT 2013 and CONSORT 2010. J Clin Epidemiol 2023;155:48-63. https://doi.org/10.1016/j.jclinepi.2023.01.003

23. Ostengaard L, Barrientos A, Boutron I, Chan AW, Collins G, Hopewell S, Moher D, Nejstgaard CH, Schulz KF, Speich B, Tang E, Tunn R, Watanabe N, Xu C, Hrobjartsson A. Development of a topic-specific bibliographic database supporting the updates of SPIRIT 2013 and CONSORT 2010. Cochrane Evid Synth Methods 2024;2:e12057. https://doi.org/10.1002/cesm.12057

24. Junqueira DR, Zorzela L, Golder S, Loke Y, Gagnier JJ, Julious SA, Li T, Mayo-Wilson E, Pham B, Phillips R, Santaguida P, Scherer RW, Gotzsche PC, Moher D, Ioannidis JP, Vohra S; CONSORT Harms Group. CONSORT Harms 2022 statement, explanation, and elaboration: updated guideline for the reporting of harms in randomized trials. J Clin Epidemiol 2023;158:149-165. https://doi.org/10.1016/j.jclinepi.2023.04.005

25. Butcher NJ, Monsour A, Mew EJ, Chan AW, Moher D, Mayo-Wilson E, Terwee CB, Chee-A-Tow A, Baba A, Gavin F, Grimshaw JM, Kelly LE, Saeed L, Thabane L, Askie L, Smith M, Farid-Kapadia M, Williamson PR, Szatmari P, Tugwell P, Golub RM, Monga S, Vohra S, Marlin S, Ungar WJ, Offringa M. Guidelines for Reporting Outcomes in Trial Reports: The CONSORT-Outcomes 2022 Extension. JAMA 2022;328:2252-2264. https://doi.org/10.1001/jama.2022.21022

26. Ghosn L, Boutron I, Ravaud P. Consolidated Standards of Reporting Trials (CONSORT) extensions covered most types of randomized controlled trials, but the potential workload for authors was high. J Clin Epidemiol 2019;113:168-175. https://doi.org/10.1016/j.jclinepi.2019.05.030

27. Boutron I, Altman DG, Moher D, Schulz KF, Ravaud P; CONSORT NPT Group. CONSORT statement for randomized trials of nonpharmacologic treatments: a 2017 update and a CONSORT extension for nonpharmacologic trial abstracts. Ann Intern Med 2017;167:40-47. https://doi.org/10.7326/M17-0046

28. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, Altman DG, Barbour V, Macdonald H, Johnston M, Lamb SE, Dixon-Woods M, McCulloch P, Wyatt JC, Chan AW, Michie S. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ 2014;348:g1687. https://doi.org/10.1136/bmj.g1687

29. Hopewell S, Chan AW, Collins GS, Hrobjartsson A, Moher D, Schulz KF, Tunn R, Aggarwal R, Berkwits M, Berlin JA, Bhandari N, Butcher NJ, Campbell MK, Chidebe RCW, Elbourne D, Farmer A, Fergusson DA, Golub RM, Goodman SN, Hoffmann TC, Ioannidis JP, Kahan BC, Knowles RL, Lamb SE, Lewis S, Loder E, Offringa M, Ravaud P, Richards DP, Rockhold FW, Schriger DL, Siegfried NL, Staniszewska S, Taylor RS, Thabane L, Torgerson D, Vohra S, White IR, Boutron I. CONSORT 2025 explanation and elaboration: updated guideline for reporting randomised trials. BMJ 2025;389:e081124. https://doi.org/10.1136/bmj-2024-081124

30. Barnes C, Boutron I, Giraudeau B, Porcher R, Altman DG, Ravaud P. Impact of an online writing aid tool for writing a randomized trial report: the COBWEB (Consort-based WEB tool) randomized controlled trial. BMC Med 2015;13:221. https://doi.org/10.1186/s12916-015-0460-y

31. Chauvin A, Ravaud P, Moher D, Schriger D, Hopewell S, Shanahan D, Alam S, Baron G, Regnaux JP, Crequit P, Martinez V, Riveros C, Le Cleach L, Recchioni A, Altman DG, Boutron I. Accuracy in detecting inadequate research reporting by early ca-

reer peer reviewers using an online CONSORT-based peer-review tool (COBPeer) versus the usual peer-review process: a cross-sectional diagnostic study. BMC Med 2019;17:205. https://doi.org/10.1186/s12916-019-1436-0

32. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R, Glanville J, Grimshaw JM, Hrobjartsson A, Lalu MM, Li T, Loder EW, Mayo-Wilson E, McDonald S, McGuinness LA, Stewart LA, Thomas J, Tricco AC, Welch VA, Whiting P, Moher D. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Syst Rev 2021;10:89. https://doi.org/10.1186/s13643-021-01626-4
33. Dimairo M, Pallmann P, Wason J, Todd S, Jaki T, Julious SA, Mander AP, Weir CJ, Koenig F, Walton MK, Nicholl JP, Coates E, Biggs K, Hamasaki T, Proschan MA, Scott JA, Ando Y, Hind D, Altman DG; ACE Consensus Group. The adaptive designs CONSORT extension (ACE) statement: a checklist with explanation and elaboration guideline for reporting randomised trials that use an adaptive design. Trials 2020;21:528. https://doi.org/10.1186/s13063-020-04334-x
34. Campbell MK, Piaggio G, Elbourne DR, Altman DG; CONSORT Group. Consort 2010 statement: extension to cluster randomised trials. BMJ 2012;345:e5661. https://doi.org/10.1136/bmj.e5661
35. Dwan K, Li T, Altman DG, Elbourne D. CONSORT 2010 statement: extension to randomised crossover trials. BMJ 2019;366:l4378. https://doi.org/10.1136/bmj.l4378
36. Yap C, Solovyeva O, de Bono J, Rekowski J, Patel D, Jaki T, Mander A, Evans TR, Peck R, Hayward KS, Hopewell S, Ursino M, Rantell KR, Calvert M, Lee S, Kightley A, Ashby D, Chan AW, Garrett-Mayer E, Isaacs JD, Golub R, Kholmanskikh O, Richards D, Boix O, Matcham J, Seymour L, Ivy SP, Marshall LV, Hommais A, Liu R, Tanaka Y, Berlin J, Espinasse A, Dimairo M, Weir CJ. Enhancing reporting quality and impact of early phase dose-finding clinical trials: CONSORT Dose-finding Extension (CONSORT-DEFINE) guidance. BMJ 2023;383:e076387. https://doi.org/10.1136/bmj-2023-076387
37. Kahan BC, Hall SS, Beller EM, Birchenall M, Chan AW, Elbourne D, Little P, Fletcher J, Golub RM, Goulao B, Hopewell S, Islam N, Zwarenstein M, Juszczak E, Montgomery AA. Reporting of factorial randomized trials: extension of the CONSORT 2010 statement. JAMA 2023;330:2106-2114. https://doi.org/10.1001/jama.2023.19793
38. Piaggio G, Elbourne DR, Pocock SJ, Evans SJ, Altman DG; CONSORT Group. Reporting of noninferiority and equivalence randomized trials: extension of the CONSORT 2010 statement. JAMA 2012;308:2594-2604. https://doi.org/10.1001/jama.2012.87802
39. Zwarenstein M, Treweek S, Gagnier JJ, Altman DG, Tunis S, Haynes B, Oxman AD, Moher D; CONSORT group; Pragmatic Trials in Healthcare (Practihc) group. Improving the reporting of pragmatic trials: an extension of the CONSORT statement. BMJ 2008;337:a2390. https://doi.org/10.1136/bmj.a2390
40. Juszczak E, Altman DG, Hopewell S, Schulz K. Reporting of multi-arm parallel-group randomized trials: extension of the CONSORT 2010 statement. JAMA 2019;321:1610-1620. https://doi.org/10.1001/jama.2019.3087
41. Vohra S, Shamseer L, Sampson M, Bukutu C, Schmid CH, Tate R, Nikles J, Zucker DR, Kravitz R, Guyatt G, Altman DG, Moher D; CENT Group. CONSORT extension for reporting N-of-1 trials (CENT) 2015 Statement. BMJ 2015;350:h1738. https://doi.org/10.1136/bmj.h1738
42. Eldridge SM, Chan CL, Campbell MJ, Bond CM, Hopewell S, Thabane L, Lancaster GA; PAFS consensus group. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. BMJ 2016;355:i5239. https://doi.org/10.1136/bmj.i5239
43. Pandis N, Chung B, Scherer RW, Elbourne D, Altman DG. CONSORT 2010 statement: extension checklist for reporting within person randomised trials. BMJ 2017;357:j2835. https://doi.org/10.1136/bmj.j2835
44. Calvert M, Blazeby J, Altman DG, Revicki DA, Moher D, Brundage MD; CONSORT PRO Group. Reporting of patient-reported outcomes in randomized trials: the CONSORT PRO extension. JAMA 2013;309:814-822. https://doi.org/10.1001/jama.2013.879
45. Manyara AM, Davies P, Stewart D, Weir CJ, Young AE, Blazeby J, Butcher NJ, Bujkiewicz S, Chan AW, Dawoud D, Offringa M, Ouwens M, Hrobjartsson A, Amstutz A, Bertolaccini L, Bruno VD, Devane D, Faria CD, Gilbert PB, Harris R, Lassere M, Marinelli L, Markham S, Powers JH 3rd, Rezaei Y, Richert L, Schwendicke F, Tereshchenko LG, Thoma A, Turan A, Worrall A, Christensen R, Collins GS, Ross JS, Taylor RS, Ciani O. Reporting of surrogate endpoints in randomized controlled trial reports (CONSORT-Surrogate): extension checklist with explanation and elaboration. BMJ 2024;386:e078524. https://doi.org/10.1136/bmj-2023-078524

46. Montgomery P, Grant S, Mayo-Wilson E, Macdonald G, Michie S, Hopewell S, Moher D; CONSORT-SPI Group. Reporting randomised trials of social and psychological interventions: the CONSORT-SPI 2018 Extension. Trials 2018;19:407. https://doi.org/10.1186/s13063-018-2733-1

47. Hopewell S, Clarke M, Moher D, Wager E, Middleton P, Altman DG, Schulz KF; CONSORT Group. CONSORT for reporting randomized controlled trials in journal and conference abstracts: explanation and elaboration. PLoS Med 2008;5:e20. https://doi.org/10.1371/journal.pmed.0050020

48. Welch VA, Norheim OF, Jull J, Cookson R, Sommerfelt H, Tugwell P; CONSORT-Equity and Boston Equity Symposium. CONSORT-Equity 2017 extension and elaboration for better reporting of health equity in randomised trials. BMJ 2017; 359:j5085. https://doi.org/10.1136/bmj.j5085

49. Hirst A, Altman DG. Are peer reviewers encouraged to use reporting guidelines?: a survey of 116 health research journals. PLoS One 2012;7:e35621. https://doi.org/10.1371/journal.pone.0035621

50. Goodman SN, Fanelli D, Ioannidis JP. What does research reproducibility mean? Sci Transl Med 2016;8:341ps12. https://doi.org/10.1126/scitranslmed.aaf5027