



JOURNAL OF Evidence-Based Practice

| e-jebp.org

Reviews

- Effectiveness and safety of intravenous lipid emulsion in women with recurrent pregnancy loss or recurrent implantation failure: a systematic review and meta-analysis
- Beyond the funnel plot: editorial and reviewer perspectives on publication bias in systematic reviews and meta-analyses in clinical medicine
- Evolution of evidence-based medicine: toward digital evidence ecosystems and artificial intelligence

Original Articles

- Sex- and age-specific associations between modifiable risk factors and incident Alzheimer's disease dementia: a nationwide cohort study in Korea
- Evaluating regional diversity in scientific communication: a comparative analysis of COVID-19 preprints and peer-reviewed publications
- Current status and challenges of Evidence-Based Medicine education in Korean medical schools: a cross-sectional survey of 21 institutions
- Variability, algorithm conformance, and accuracy of large language model-based tools for risk-of-bias (RoB 2.0) assessment of randomized trials: a pilot study

- Interpreting patient values and preferences in clinical practice guideline development: distinguishing reliance on physician judgment from treatment outcome preferences – an exploratory multicenter study in patients with anal cancer

Case Reports

- General anesthesia with remimazolam in a patient with leigh syndrome: a case report
- Successful endoscopic internal drainage for infected intramural esophageal dissection complicated by mediastinal abscess: a case report



JOURNAL OF

<https://e-jebp.org>

eISSN 3059-1996

Evidence-Based Practice

Vol. 2 No. 2, September 2026

Aims & Scope

Journal of Evidence-Based Practice (J Evid-Based Pract, JEBP) aims to present 1) Original evidence-based research on important issues in healthcare, 2) Methods, tools, and concepts essential for evidence-based medicine (EBM), education and practice, 3) Perspectives, debates, analyses, and opinions on reliable evidence and related topics in evidence-based medicine.

Open Access

Articles published in JEBP are open-access, distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0>), which permits unrestricted non-commercial use, distribution, and the reproduction in any medium, provided that the original work is properly cited.

Publisher: Korean Society of Evidence-Based Medicine

Editor-in-Chief: Hyun Kang, MD, PhD

Editorial Office

Korean Society of Evidence-Based Medicine

Room 219, 2nd floor of Apex Buliding, 30, bamgogae-ro 1-gil, Gangnamgu, Seoul 06349, Korea

Tel: +82-2-459-8206 Fax: +82-2-459-8256 E-mail: ksebmoffice@ksebm.or.kr

Printing Office

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

Editorial Board

Editor-in-Chief

Hyun Kang

Department of Anesthesiology and Pain Medicine, Chung-Ang University, Korea

Deputy Editor

Oh Haeng Lee

Department of Anesthesiology and Pain Medicine, Chung-Ang University, Korea

Editorial Boards

Eun Jin Ahn

Department of Anesthesiology and Pain Medicine, Chung-Ang University, Korea

Moon Seong Baek

Department of Internal Medicine, Chung-Ang University, Korea

Eun Joo Choi

Department of Anesthesiology and Pain Medicine, Seoul National University, Korea

Jong Han Choi

Department of Internal Medicine, Konkuk University, Korea

Hyun Jung Kim

Department of Preventive Medicine, Korea University, Korea

Soo Young Kim

Department of Family Medicine, Hallym University, Korea

Su Hyun Kim

Department of Internal Medicine, Chung-Ang University, Korea

Bada Yang

Department of Epidemiology and Health Economics, University Medical Center Utrecht, Netherlands

Ethics Editor

Sang-Il Lee

University of Ulsan, Korea

Statistical Editors

Soyeon Ahn

Seoul National University, Korea

Sang Kyu Kwak

Daegu Catholic University, Korea

Reviews

- 35 Effectiveness and safety of intravenous lipid emulsion in women with recurrent pregnancy loss or recurrent implantation failure: a systematic review and meta-analysis
Jungeun Park, Jinyoung Chang, Haine Lee, Bo Hyon Yun, Dong Ah Park
- 45 Beyond the funnel plot: editorial and reviewer perspectives on publication bias in systematic reviews and meta-analyses in clinical medicine
Hyun Kang
- 56 Evolution of evidence-based medicine: toward digital evidence ecosystems and artificial intelligence
Soo Young Kim

Original Articles

- 62 Sex- and age-specific associations between modifiable risk factors and incident Alzheimer's disease dementia: a nationwide cohort study in Korea
Hyun Jung Kim, Mookyoung Oh
- 73 Evaluating regional diversity in scientific communication: a comparative analysis of COVID-19 preprints and peer-reviewed publications
Dong Hyun Kim, Kyulee Jeon, Seng Chan You
- 84 Current status and challenges of Evidence-Based Medicine education in Korean medical schools: a cross-sectional survey of 21 institutions
Hyun Jung Kim, Mookyoung Oh, Jae Hung Jung, Geun Joo Choi, Heather Swan, Soo Young Kim
- 91 Variability, algorithm conformance, and accuracy of large language model-based tools for risk-of-bias (RoB 2.0) assessment of randomized trials: a pilot study
Jiae Choi, Heather Swan, Min Jung Kim, Hyun Jung Kim
- 98 Interpreting patient values and preferences in clinical practice guideline development: distinguishing reliance on physician judgment from treatment outcome preferences – an exploratory multicenter study in patients with anal cancer
Hyo Seon Ryu, Jung-Myun Kwak, Kil-Yong Lee, Dalyong Kim, Dong Hyun Kang, Heather Swan, Hyun Jung Kim

Case Reports

- 107 General anesthesia with remimazolam in a patient with leigh syndrome: a case report
Young Ki Kim, Dong-Hyun Kang
- 112 Successful endoscopic internal drainage for infected intramural esophageal dissection complicated by mediastinal abscess: a case report
Jae Yong Park, Byung Joon Park, Beom Jin Kim, Jae Gyu Kim

Effectiveness and safety of intravenous lipid emulsion in women with recurrent pregnancy loss or recurrent implantation failure: a systematic review and meta-analysis

Jungeun Park¹, Jinyoung Chang¹, Haine Lee¹, Bo Hyon Yun², Dong Ah Park¹

¹National Evidence-based Healthcare Collaborating Agency (NECA), Seoul, Korea

²Department of Obstetrics and Gynecology, Severance Hospital, Yonsei University College of Medicine, Seoul, Korea

Intralipid, an intravenous lipid emulsion, is widely used for women with recurrent pregnancy loss (RPL) or recurrent implantation failure (RIF), despite limited evidence. This systematic review evaluated the effectiveness and safety of intralipid in this population. We searched seven electronic databases from inception to April 14, 2025. Randomized controlled trials (RCTs) and non-randomized comparative studies (NRCSs) were included for effectiveness analysis; case series and reports for safety. Two reviewers independently screened and extracted data. Risk of bias was assessed using RoB 2.0 and RoBANS 2.0. Meta-analyses used a random-effects model (risk ratios [RR] with 95% confidence intervals [CI]). Evidence certainty was assessed using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach. Seventeen studies were included (8 effectiveness, 11 safety). Pooled RCT data showed intralipid significantly improved clinical pregnancy rate vs. no treatment (RR 2.31, 95% CI 1.42–3.74; $I^2 = 0\%$; GRADE certainty: low); this effect was not observed in the pooled NRCS data for live birth rate or clinical pregnancy rate, and one large NRCS reported a significantly higher miscarriage rate in the intralipid group (RR 1.12, 95% CI 1.04–1.20). No significant differences were observed vs. intravenous immunoglobulin (IVIG) or steroids. Serious adverse events were rare. GRADE certainty was very low to low across all outcomes. Current evidence does not support the routine use of intralipid in women with RPL or RIF. While pooled RCT data suggest a potential benefit in clinical pregnancy rate, evidence certainty is low, and this benefit was not observed in the pooled NRCS data. Adequately powered RCTs with standardized protocols and live birth rate as the primary endpoint are needed.

Keywords: Intralipid; Recurrent implantation failure; Recurrent pregnancy loss; Systematic review; Meta-analysis; Immunotherapy

Introduction

Infertility is a major and growing reproductive health concern worldwide. In South Korea, where advanced maternal age and ultra-low birth rates are increasingly prevalent, infertility carries particular public health significance. According to data from the Health Insurance Review and Assessment

Service (HIRA), the number of patients diagnosed with infertility in South Korea reached approximately 289,000 in 2024, representing a 1.3-fold increase from 228,000 in 2020 [1]. The number of individuals undergoing assisted reproductive technology (ART) procedures similarly increased 1.2-fold, from 130,000 in 2020 to 161,000 in 2024. Repeated ART failure imposes substantial psychological and financial burdens on

Received: August 1, 2026; **Revised:** August 18, 2026; **Accepted:** August 25, 2026

Corresponding author: Dong Ah Park

Email: pda124@neca.re.kr

Bo Hyon Yun

Email: garfieldzz@yuhs.ac

© 2026 Korean Society of Evidence-Based Medicine

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

affected women and their families, prompting exploration of alternative treatment strategies.

Immunological factors have been implicated in implantation failure and pregnancy loss, and various immunomodulatory therapies have been introduced into clinical practice. Among these, Intralipid, an intravenous lipid emulsion originally approved for total parenteral nutrition (TPN) and holds no regulatory indication for infertility treatment or immunomodulation. Nevertheless, several basic and observational studies have reported that intralipid may inhibit natural killer (NK) cell cytotoxicity [2,3], reduce Th1/Th2 immune activation [4,5], and improve the implantation environment. However, most of these studies were limited by small sample sizes, lack of prospective controls, and inconsistent diagnostic criteria. High-quality evidence confirming the effectiveness of intralipid in infertile patients remains insufficient.

Kumar et al. noted that intralipid may be considered in certain immunological high-risk groups but is not supported by sufficient evidence to be recommended as standard care for women with recurrent implantation failure (RIF) or recurrent pregnancy loss (RPL) [6]. The Human Fertilisation and Embryology Authority (HFEA) classifies intralipid and other immunomodulatory treatments as treatments with "insufficient evidence and potential harm" and does not recommend their use for fertility treatment [7]. Despite concerns raised by international regulatory and professional bodies regarding the limited evidence base for intralipid in fertility treatment, intralipid infusions are widely administered in South Korea in a non-reimbursed setting, including repeated infusions justified solely by marginal elevations in NK cell activity, raising concerns about the clinical expansion of insufficiently evidenced therapies.

This systematic review was conducted to generate evidence on the clinical effectiveness and safety of intralipid infusion in women with RPL, RIF and repeated IVF failure, with the aim of supporting appropriate and rational use of intralipid in infertility treatment.

Methods

This systematic review and meta-analysis was prospectively registered in PROSPERO (CRD420251039275). The review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines [8].

Eligibility criteria

Eligible studies were those enrolling women with RPL or RIF (including studies defining RIF as two or more failed IVF cycles) as the target population, investigating intralipid infusion

as the intervention for pregnancy purposes, and comparing outcomes with placebo, no treatment, or other immunomodulatory treatments (including IVIG or steroids). For the purposes of this review, repeated IVF failure was considered to overlap clinically with RPL and RIF and was therefore included within the eligible population. Eligible study designs for effectiveness analysis included randomized controlled trials (RCTs) and non-randomized comparative studies (NRCs) (cohort and case-control studies) with at least 10 participants per arm. Studies reporting only safety outcomes (case series and case reports) were also included in the safety analysis. Exclusion criteria included animal studies, pre-clinical studies, non-original research (systematic reviews, narrative reviews, guidelines, editorials, opinions), grey literature, theses, and retracted publications. Conference abstracts were excluded from the effectiveness analysis because of insufficient methodological and outcome information but were eligible for the safety analysis when they provided identifiable adverse-event data.

Outcomes

The primary outcomes were live birth rate and clinical pregnancy rate. Clinical pregnancy was defined as the presence of a gestational sac with fetal cardiac activity confirmed by ultrasonography; where studies used alternative definitions, reported values were extracted as documented and noted accordingly. The secondary outcome was miscarriage rate, defined as pregnancy loss prior to 20 weeks of gestation. Safety outcomes included maternal adverse reactions, obstetric complications, and neonatal outcomes such as congenital anomalies.

Literature search

We searched five major international electronic databases—Ovid-MEDLINE, Ovid-EMBASE, Cochrane Central Register of Controlled Trials (CENTRAL), Web of Science, and Scopus—as well as two Korean databases, KoreaMed and KMBASE. There were no restrictions on publication year, and the final search date was April 14, 2025. Search strategies combined Medical Subject Headings (MeSH) terms and free-text terms related to infertility, ART, and intralipid, covering the population (P) and intervention (I) components. Detailed search strategies are provided in the [Supplementary Methods S1](#).

Study selection and data extraction

Two reviewers independently screened titles and abstracts, then full texts, according to the predefined eligibility criteria. Disagreements were resolved by discussion with a third re-

viewer. Data extraction covered study characteristics (country, year, design, participating institutions, study period), participant characteristics (inclusion/exclusion criteria, number of participants, age, clinical group), intervention characteristics (intralipid formulation, dose, timing, frequency, concurrent treatments, prior NK cell testing), and outcome data (outcome measure, measurement method, reported values, summary statistics, timing of assessment). For studies reporting both cycle-level and woman-level data, woman-level data were preferentially extracted for the intralipid versus no-treatment comparison to minimize unit-of-analysis concerns arising from non-independent repeated cycles. For example, although Mrosk et al. reported most outcomes at the treatment-cycle level, the intralipid versus no-treatment comparison was based on a woman-level subgroup (patients with elevated uterine natural killer cell counts) and was extracted accordingly.

Risk of bias assessment

Risk of bias was assessed independently by two reviewers, with discrepancies resolved by a third reviewer. RCTs were assessed using the Cochrane Risk of Bias 2.0 (RoB 2.0) tool across five domains. NRCs were assessed using the RoBANS ver 2.0 tool. Risk of bias assessment was not performed for case series and case reports included for safety analyses only.

Data synthesis and statistical analysis

When data could be pooled quantitatively, meta-analysis was performed; otherwise, a qualitative synthesis was conducted. Given anticipated clinical heterogeneity across studies, a random-effects model was applied throughout. For dichotomous outcomes, the Mantel-Haenszel method was used to calculate RR and 95% CI. For continuous outcomes, weighted mean difference (WMD) or standardized mean difference (SMD) with 95% CI was used depending on whether outcomes were measured on the same or different scales. Heterogeneity was assessed visually using forest plots and statistically using Cochrane's Q statistic ($P < 0.10$) and I^2 statistics ($\geq 50\%$) [9]. Subgroup analyses were prespecified by patient group (RIF or RPL), sample size, and intervention dose. Sensitivity analyses were conducted according to patient group (RIF only), NK cell testing status, risk of bias level, and whether adjusted estimates were reported. All meta-analyses were performed using RevMan 5.4. Statistical significance was set at $P < 0.05$. Funnel plots and Egger's test for small-study effects were planned to assess reporting bias when at least 10 studies were available for a given meta-analysis. As no synthesis included 10 or more studies, these assessments were not performed.

Certainty of evidence

The certainty of evidence was assessed for direct intervention effect outcomes using the GRADE approach [10], considering risk of bias, inconsistency, indirectness, imprecision, and publication bias.

Results

Literature search and study selection

A total of 3,733 records were identified through searches of seven electronic databases. After removal of 1,398 duplicates, 2,335 records underwent title and abstract screening, of which 2,246 were excluded. The remaining 89 reports were retrieved for full-text assessment, and 17 studies were ultimately included: 8 for effectiveness analysis (3 RCTs, 5 NRCs) and 11 for safety analysis (4 RCTs, 2 NRCs, 4 case series, 1 case report). The study selection process is shown in the PRISMA flow diagram (Fig. 1), and reasons for exclusion of full-text-reviewed studies are presented in [Supplementary Table S1](#).

Characteristics of included studies for effectiveness analysis

The three RCTs for effectiveness analysis were published in 2016 and 2022, conducted in Russia, Iran, and China [11–13], and enrolled women with RIF (2 studies) or women undergoing IVF (1 study). Two RCTs compared intralipid with no treatment/placebo [12,13], and one compared intralipid with IVIG [11]. The five NRCs [14–18] were all retrospective cohort studies, published between 2018 and 2024, and conducted in the United Kingdom ($n = 1$), France ($n = 2$), Germany ($n = 1$), and the United States ($n = 1$). Target populations included RIF (2 studies), a mix of RIF and RPL or IVF recipients (3 studies). Comparators were no treatment [15,17,18], prednisolone [14], and both no treatment and steroids [16]. Across studies, 20% intralipid was used in most studies. Doses ranged from 100 to 250 mL per infusion. The number of infusions ranged from 1 to 5, with varying timing relative to embryo transfer. Characteristics of included studies and details of intervention are summarized in [Table 1](#).

Risk of bias assessment

All three RCTs reported adequate randomization but failed to report allocation concealment, resulting in “some concerns” in the randomization process domain. Patient blinding was absent or unclear, and deviation from intended intervention was rated as “some concerns” across all three. One RCT [11] was rated “high risk” overall due to 40% loss to follow-up for live birth outcomes. The remaining two RCTs [12,13] were

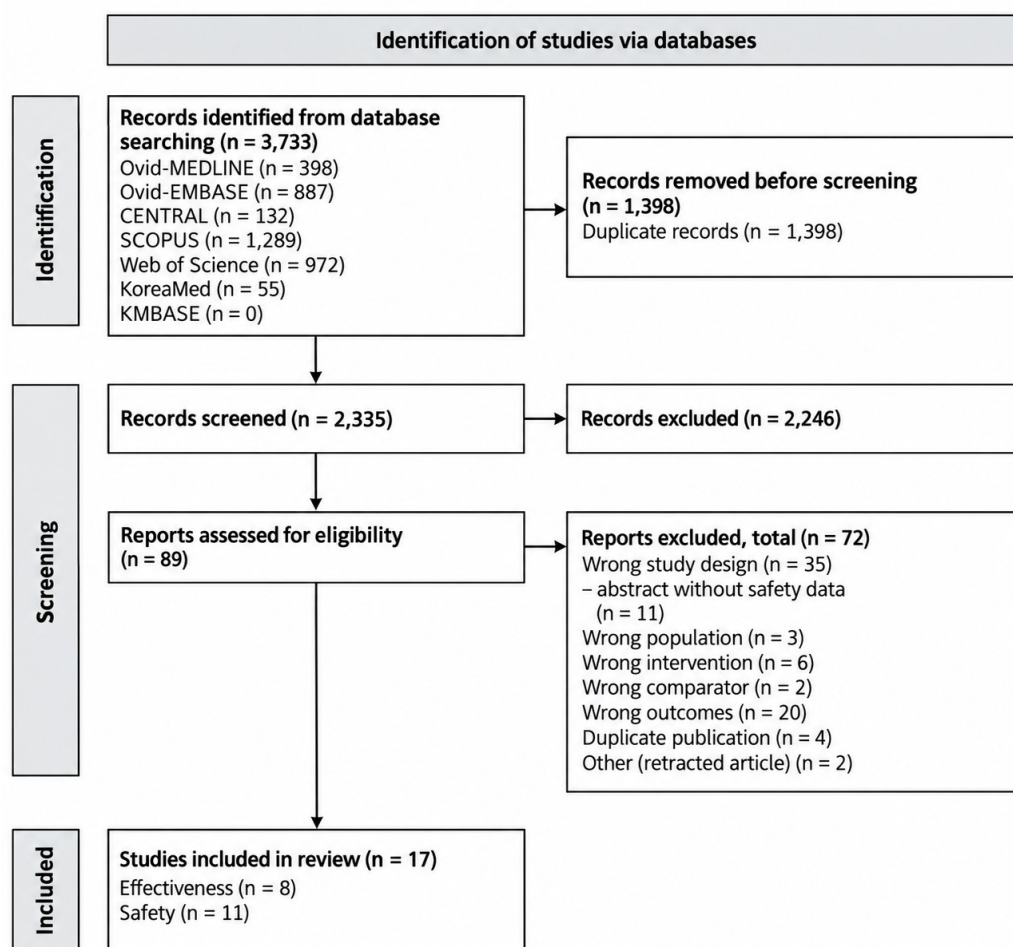


Fig. 1. PRISMA flow diagram of study selection.

rated “some concerns” overall (Supplementary Figure S1).

The five NRCS were assessed using the RoBANS ver 2.0 tool across eight domains. The main sources of bias were comparability of groups, selection of study participants, and confounding variables. Martini et al. was rated “high risk” in both the comparability and confounding domains, as it used a historical RCT control group with differing eligibility criteria and baseline characteristics, and applied only unadjusted statistical tests without controlling for confounders [14]. Most remaining studies were rated “unclear” for participant selection due to their retrospective designs, in which outcomes were already known at enrollment. Blinding of outcome assessors and selective outcome reporting were rated “low risk” across all studies, as pregnancy outcomes are objectively ascertained and primary outcomes were consistently reported (Supplementary Figure S2).

Effectiveness: intralipid versus no treatment

No RCT reported live birth rate. Among three NRCS ($n =$

21,320), there was no statistically significant difference in live birth rate between intralipid and no-treatment groups (RR 2.10, 95% CI 0.89–4.96; $I^2 = 63\%$) (Fig. 2).

Two RCTs showed a statistically significant improvement in clinical pregnancy rate in the intralipid group compared with no treatment (RR 2.31, 95% CI 1.42–3.74; $I^2 = 0\%$) (Fig. 3). In contrast, pooled results from four NRCS did not show a statistically significant difference in clinical pregnancy rate (RR 1.84, 95% CI 0.95–3.59; $I^2 = 57\%$) (Fig. 4). One NRCS [15] reported a statistically significant increase in miscarriage rate in the intralipid group ($n = 20,315$; RR 1.12, 95% CI 1.04–1.20). These results are summarized in Table 2.

In RCT-based subgroup analyses by patient group, the single study enrolling a mixed RIF and repeated IVF failure population showed a significantly higher clinical pregnancy rate (RR 2.15, 95% CI 1.25–3.70) [13]. In NRCS-based subgroup analyses, pooled results from two studies enrolling mixed RIF/RPL populations also showed a significantly higher live birth rate (RR 3.60, 95% CI 1.51–8.57; $I^2 = 0\%$). Prespecified

Table 1. Characteristics of Intralipid Infusion Regimens Across Included Studies

No.	Study	Country/ Design	Population	N (intervention/control)	Intervention	Comparator
1	Bespalova 2022 [12]	Russia, RCT	IVF failure (≥2 cycles)	97 (41/56)	Intralipid 20%, 200 mL x 3, once monthly	No treatment
2	Peivandi 2022 [13]	Iran, RCT	RIF (≥2 failures)	80 (40/40)	Intralipid 20%, 250 mL(diluted with 2 mL) x 1, 2 days before ET	Normal saline 250 mL x 1, 2 days (2 h) before ET
3	Meng 2016 [11]	China, RCT	RIF (≥3 failures)	154 (76/78)	Intralipid 20%, 250 mL x ≥3, cycle day 3; pre-preg q2wk; post-preg weekly (≤12 wk)	IVIIG 25 g, cycle day 8–10; pre-preg monthly; post-preg weekly (≤12 wk)
4	Khairy 2024 [15]	UK, retrospective cohort	RIF	12,949 (1,960/10,989)	Intralipid 20%, 100 mL x 1, cycle day 4–9	No treatment (standard ART)
5	Mrosk 2024 [18]	Germany, retrospective cohort	Women undergoing IVF/ICSI with ES treatment (elevated uNK cell counts)	79 (67/12)	Intralipid 20%, 150 mL x 3; preceding cycle, oocyte retrieval day, and if pregnant	No treatment
6	Kolanska 2021 [16]	France, retrospective cohort	RIF (≥3 failures)	60 (18/20/22)	Intralipid x 4, day 8 after ET + 3, 5, 9 wk of pregnancy	(1) Corticosteroid 10 mg/day, (3 wk before transfer ~ 9 wk of amenorrhea)
7	Placais 2020 [17]	France, retrospective cohort	RPL or RIF (≥3)	62 (26/36)	Intralipid x 5, day 8 after ET+ 3, 5, 9 wk of pregnancy	(2) No treatment
8	Martini 2018 [14]	USA, retrospective cohort	RPL or RIF	147 (127/20)	Intralipid 20%, 250 mL (diluted with 4 mL) x 3, 7–10 d before ET/fertilization + 6 wk of pregnancy + 10 wk of pregnancy	Prednisolone: 20 mg x 6 wk, 10 mg x 1 wk, 5 mg x 1 wk

ART, assisted reproductive technology; ES, endometrial scratching; ET, embryo transfer; IVF, in vitro fertilization; ICSI, intracytoplasmic sperm injection; IVIG, intravenous immunoglobulin; PSM, propensity score matching; RCT, randomized controlled trial; RIF, recurrent implantation failure; RPL, recurrent pregnancy loss.

Table 2. GRADE Summary of Findings: Intralipid Infusion versus No Treatment

Participants (studies) Follow-up	Overall certainty of evidence	Study event rates (%)		Relative effect (95% CI)	Anticipated absolute effects	
		With No Tx	With intralipid		Risk with No Tx	Risk difference with intralipid
Live birth rate 21,320 (3 NRCS)	⊕○○○ Very low ^{a,b}	2,648/16,767 (15.8%)	963/4,553 (21.2%)	RR 2.10 (0.89 to 4.96)	2,648/16,767 (15.8%)	174 more per 1,000 (from 17 fewer to 625 more)
Clinical pregnancy rate 159 (2 RCTs)	⊕⊕○○ Low ^{cd}	17/87 (19.5%)	31/72 (43.1%)	RR 2.31 (1.42 to 3.74)	17/87 (19.5%)	256 more per 1,000 (from 82 more to 535 more)
21,330 (4 NRCS)	⊕○○○ Very low ^{a,b}	3,651/16,769 (21.8%)	1,260/4,561 (27.6%)	RR 1.84 (0.95 to 3.59)	3,651/16,769 (21.8%)	183 more per 1,000 (from 11 fewer to 564 more)
Miscarriage rate 20,315 (1 NRCS)	⊕○○○ Very low ^a	2,638/15,912 (16.6%)	818/4,403 (18.6%)	RR 1.12 (1.04 to 1.20)	2,638/15,912 (16.6%)	20 more per 1,000 (from 7 more to 33 more)

* Statistically significant (P < 0.05).

CI, confidence interval; GRADE, Grading of Recommendations Assessment, Development and Evaluation; NRCS, non-randomized comparative studies; RCT, randomized controlled trial; RR, risk ratio.

^aDowngraded by 1 level for risk of bias, owing to limitations in domains such as comparability of groups and selection of participants. ^bDowngraded by 2 levels for imprecision, as the confidence interval crosses both no effect and appreciable benefit. ^cDowngraded by 1 level for risk of bias, owing to serious concerns in 2 or more domains (e.g., the randomization process and deviations from intended interventions). ^dDowngraded by 1 level for imprecision, as the optimal information size (OIS) was not met (48 events).

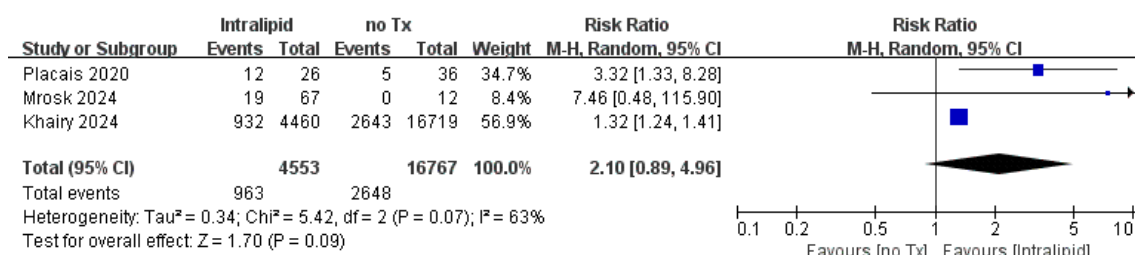


Fig. 2. Forest plot of live birth rate (intralipid vs. no treatment, NRCS).

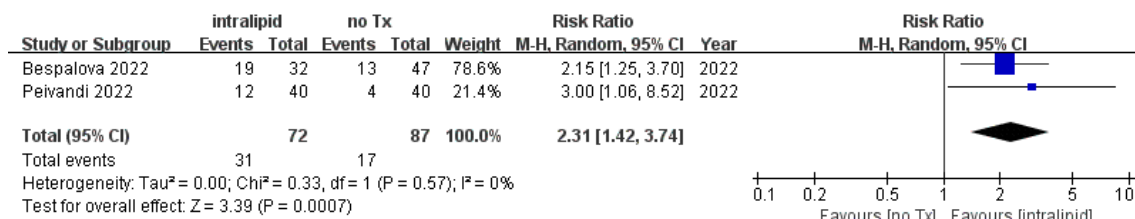


Fig. 3. Forest plot of clinical pregnancy rate (intralipid vs. no treatment, RCT).

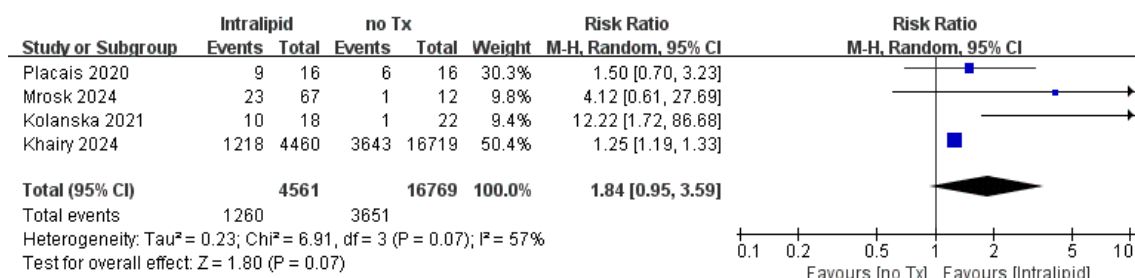


Fig. 4. Forest plot of clinical pregnancy rate (intralipid vs. no treatment, NRCS).

subgroup analyses by sample size and intralipid dose could not be performed due to an insufficient number of studies within each category or inconsistent reporting of dose information across studies (Supplementary Table S2).

Sensitivity analyses

Among RCTs restricted to RIF women, clinical pregnancy rate remained significantly higher in the intralipid group (1 RCT; RR 3.00, 95% CI 1.06–8.52). Among NRCSs restricted to RIF women, live birth rate was significantly higher (1 NRCS; RR 1.32, 95% CI 1.24–1.41), and miscarriage rate was also significantly higher (1 NRCS; RR 1.12, 95% CI 1.04–1.20), while clinical pregnancy rate was not significant (2 NRCSs; RR 1.29, 95% CI 0.64–2.61; $I^2 = 81\%$). Among NRCSs using NK cell testing as a patient selection criterion [15,18], neither live birth rate (RR 1.79, 95% CI 0.49–6.47; $I^2 = 35\%$) nor clinical pregnancy rate (RR 1.53, 95% CI 0.64–3.65; $I^2 = 33\%$) showed significant differences. In the single NRCS rated as low risk

of bias [16], clinical pregnancy rate showed a large and significant effect (RR 12.22, 95% CI 1.72–86.68). Although the adjusted estimate for clinical pregnancy rate from the single study reporting adjusted results (Kolanska 2021; adjusted OR 36.04, 95% CI 2.26–575.18) nominally favored intralipid, the confidence interval was extremely wide, rendering this estimate highly imprecise and unreliable. Overall, findings were consistent with the main analyses and did not alter their interpretation: effect estimates from single-study analyses were imprecise, and the one available adjusted estimate carried an extremely wide confidence interval, limiting its reliability (Supplementary Table S2).

Effectiveness: intralipid versus IVIG or steroids

When compared with other immunomodulatory treatments, intralipid demonstrated no significant advantage over either IVIG or steroids. In one RCT [11] ($n = 154$) comparing intralipid with IVIG, no significant differences were observed in

live birth rate (RR 0.83, 95% CI 0.63–1.10), clinical pregnancy rate (RR 1.07, 95% CI 0.96–1.20), or miscarriage rate (RR 0.63, 95% CI 0.28–1.44). Similarly, pooled analysis of two NRSs ($n = 185$) comparing intralipid with steroid-based treatment (prednisolone or steroids) revealed no statistically significant differences in live birth rate (Martini et al; RR 0.93, 95% CI 0.52–1.66) or clinical pregnancy rate (Martini et al and Kolarska et al; RR 0.84, 95% CI 0.57–1.24; $I^2 = 33\%$) ([Supplementary Table S3](#)).

GRADE certainty of evidence

For the comparison of intralipid versus no treatment, the certainty of evidence was “low” for clinical pregnancy rate based on RCT data (downgraded for risk of bias and imprecision), and “very low” for live birth rate ([Table 2](#)), clinical pregnancy rate from NRCs, and miscarriage rate. For comparisons with IVIG and steroids, all outcomes were rated “very low” due to risk of bias and imprecision ([Supplementary Table S3](#)).

Safety findings

Safety data were available from 4 RCTs [[11,13,19,20](#)], 2 NRCs [[21,22](#)], and 5 case studies [[23–27](#)].

Of the RCTs, no adverse events were reported in the intralipid group in three studies [[11,13,19](#)]; comparator: no treatment or IVIG), while one study reported congenital external ear malformation in two cases receiving intralipid [[20](#)]. In the control groups, no adverse events were reported except in one RCT comparing intralipid with IVIG, in which mild fever ($n = 5$) and headache ($n = 2$) occurred in the IVIG group [[11](#)].

Among the two NRCs, findings were contrasting. No adverse events were reported in the study comparing intralipid with no treatment [[21](#)]. In contrast, the study comparing intralipid with low molecular weight heparin (LMWH) reported obstetric complications in the intralipid group, including pre-eclampsia ($n = 1$), intrauterine growth restriction ($n = 1$), and placental abruption ($n = 1$), although none were classified as serious adverse events; no serious adverse events were reported in the LMWH group [[22](#)].

Of the five case studies and case series, no adverse events were reported in four reports [[23,24,26,27](#)]. One report, involving patients with elevated NK cell levels treated with intralipid, described facial flushing with pre-seizure symptoms ($n = 1$) and asymmetric fetal intrauterine growth restriction ($n = 1$) [[25](#)]. Full details of reported cases are presented in [Supplementary Table S4](#).

Discussion

This systematic review and meta-analysis evaluated the effec-

tiveness and safety of intralipid infusion in women with RIF or RPL. Of the two randomized controlled trials comparing intralipid with no treatment in women with RIF or RPL, neither reported data on miscarriage rate. Nevertheless, pooled data from these two trials demonstrated a significantly higher clinical pregnancy rate in the intralipid group compared with the no-treatment group (RR 2.31, 95% CI 1.42–3.74; GRADE certainty: low); however, this effect was not observed in the pooled NRCs data. One large observational study additionally reported a significantly increased miscarriage rate in the intralipid group. When compared with other immunomodulatory treatments such as IVIG or steroids, no statistically significant differences were detected, although the estimates were imprecise and the certainty of evidence was very low. No severe or life-threatening adverse events directly attributable to intralipid were reported, although safety data were limited and incompletely ascertained. In a small number of studies, congenital anomalies, obstetric complications, and fetal growth restriction were reported in the intralipid group; however, a causal relationship with intralipid could not be established given the limited ascertainment and reporting of adverse events across studies. Long-term follow-up data on maternal and neonatal outcomes were also absent across all included studies.

Nevertheless, our findings are broadly consistent with prior systematic reviews, which have similarly concluded that the evidence is insufficient to support or refute the use of intralipid in women with RPL or RIF [[6,28–31](#)]. The present review differs from these prior reviews in several key respects, offering an updated and more rigorous synthesis of the current evidence. First, this review incorporates recently published studies not included in earlier meta-analyses, while excluding studies whose findings can no longer be considered valid. Al-Zebeidi et al. [[32](#)], cited in five prior systematic reviews, and Dakhly et al. [[33](#)], cited in four prior systematic reviews, were excluded after confirming their retraction from peer-reviewed journals in September 2025 and May 2024, respectively, thereby reducing potential distortion of the pooled estimates. Second, we made explicit efforts to minimize clinical and methodological heterogeneity by analyzing effectiveness data separately by study design (RCT vs. NRCs) and by indication (RIF vs. RPL). This approach revealed discordant effect estimates between randomized and non-randomized evidence and highlighted the absence of RCT evidence in RPL-only populations. Despite the resulting reduction in per-analysis sample size, a potential treatment effect could still be discerned in several RCT-based analyses, even though the certainty of evidence remained low. Third, this review provides a more comprehensive assessment of

safety by additionally incorporating case reports and conference abstracts, sources often excluded from prior systematic reviews, thereby capturing a broader range of real-world safety signals than previously reported.

The biological rationale for intralipid in fertility treatment rests primarily on its proposed ability to suppress peripheral and uterine NK cell activity, thereby improving the immune tolerance required for embryo implantation. However, the clinical evidence for NK cell-guided patient selection remains weak. In our sensitivity analyses restricted to studies using NK cell testing as a selection criterion, no significant benefit was demonstrated for either live birth or clinical pregnancy rate. Furthermore, the two included studies differed in the type of NK cell testing used as the selection criterion: one assessed peripheral blood NK cell proportion and activity [15], while the other assessed uterine NK cell density measured immunohistochemically in endometrial biopsy specimens [18]. Given that peripheral and uterine NK cell measurements reflect distinct compartments and lack a validated correlation with each other, this heterogeneity in testing modality, in addition to differing definitions of NK cell abnormality across studies, further limits the interpretability of these findings.

This review has several limitations. First, the certainty of evidence was very low to low across all outcomes, primarily due to risk of bias in the primary studies, substantial heterogeneity, and imprecision. Well-designed and rigorously conducted clinical trials are needed to establish the clinical effectiveness of intralipid infusion therapy. Second, most included studies were small and conducted at single centers in limited geographic settings (Europe, Middle East, East Asia); no studies were conducted in Korean patients, despite widespread empirical use of intralipid in South Korea. Third, although this review's inclusion of conference abstracts and case reports broadened the safety evidence base, these sources inherently lack the methodological rigor of peer-reviewed studies, restricting quantitative synthesis and precluding firm conclusions regarding the safety profile of intralipid. Further well-designed studies with systematic safety reporting are needed to determine whether the observed adverse findings, including congenital anomalies and obstetric complications, reflect a true potential harm associated with intralipid or occurred independently of treatment. Fourth, publication bias was not formally assessed due to the small number of eligible studies. Finally, this review did not fully differentiate treatment effects according to the underlying indication; in particular, whether the effectiveness of intralipid differs between women with RIF, the primary indication for its use, and those with RPL remains to be clarified. Future studies stratified by

indication are warranted to determine whether specific subgroups derive greater benefit from intralipid therapy.

In conclusion, current evidence is insufficient to support the clinical safety and effectiveness of intralipid infusion therapy in women with RIF or RPL. Given these gaps, intralipid should not be used indiscriminately as a routine empirical treatment outside the context of well-designed clinical trials. Future research should prioritize adequately powered RCTs with standardized treatment protocols—including defined dose, timing, and indication-specific patient selection criteria—using live birth rate as the primary endpoint, together with systematic long-term safety monitoring for both mother and offspring.

Conflict of Interest

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

Funding

This study was supported by the National Evidence-based Healthcare Collaborating Agency (NECA), Republic of Korea (project number: NECA-A-25-009). The funders had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Data Availability Statement

The datasets generated and/or analyzed in the current study are available from the corresponding author on reasonable request.

Ethics Approval and Consent to Participate

Not applicable.

Authors' Contributions

Conception and design: DAP, BHY. Administrative support: JYC, HIL. Provision of study materials: JEP, JYC, DAP. Collection and assembly of data: JEP, JYC, HIL. Data analysis and interpretation: JEP, DAP, BHY. Manuscript writing: JEP, DAP, BHY. Final approval of manuscript: All authors.

Acknowledgments

The authors would like to express their sincere gratitude to Prof. Joon Cheol Park, Prof. Aera Han, and Prof. Seul Ki Kim

for their valuable clinical advice and expert consultation throughout this study.

Disclosure of AI use

In preparing this manuscript, the authors used Claude Sonnet 4.6 (Anthropic, San Francisco, CA, USA) to assist with Korean–English translation and editorial refinement of the text. All AI-generated content was critically reviewed, revised, and verified by the authors, who take full responsibility for the accuracy and integrity of the work.

Supplementary Materials

Supplementary files are available from <https://doi.org/10.63528/jebp.2026.00009>

ORCID

Jungeun Park, <https://orcid.org/0000-0002-1129-5495>
Jinyoung Chang, <https://orcid.org/0000-0002-0441-0061>
Haine Lee, <https://orcid.org/0000-0002-9643-1176>
Bo Hyon Yun, <https://orcid.org/0000-0001-5703-797X>
Dong Ah Park, <https://orcid.org/0000-0001-7225-3152>

References

1. Health Insurance Review and Assessment Service (HIRA). National health insurance statistics on infertility treatment procedures (assisted reproductive technology) [Internet]. Wonju: HIRA; 2025 [cited 2025 Dec 5]. Available from <https://openda-ta.hira.or.kr/op/opc/olapMfrnIntrsInnsInfoTab1.do>
2. Coulam CB, Roussev RG. Correlation of NK cell activation and inhibition markers with NK cytotoxicity among women experiencing immunologic implantation failure after in vitro fertilization and embryo transfer. *J Assist Reprod Genet* 2003; 20: 58-62.
3. Beaman KD, Ntrivalas E, Mallers TM, Jaiswal MK, Kwak-Kim J, Gilman-Sachs A. Immune etiology of recurrent pregnancy loss and its diagnosis. *Am J Reprod Immunol* 2012; 67: 319-25.
4. Moffett A, Shreeve N. First do no harm: uterine natural killer (NK) cells in assisted reproduction. *Hum Reprod* 2015; 30: 1519-25.
5. Check JH, Check D. Intravenous intralipid therapy is not beneficial in having a live delivery in women aged 40-42 years with a previous history of miscarriage or failure to conceive despite embryo transfer undergoing in vitro fertilization-embryo transfer. *Clin Exp Obstet Gynecol* 2016; 43: 14-5.
6. Kumar P, Marron K, Harriy C. Intralipid therapy and adverse reproductive outcome: is there any evidence? *Reprod Fertil* 2021; 2: 173-86.
7. Human Fertilisation and Embryology Authority. Immunological tests and treatments for fertility [Internet]. London: HFEA; 2023 [cited 2025 Apr 14]. Available from <https://www.hfea.gov.uk/treatments/treatment-add-ons/immunological-tests-and-treatments-for-fertility/>
8. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021; 372: n71.
9. Higgins JP. Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified. *Int J Epidemiol* 2008; 37: 1158-60.
10. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008; 336: 924-6.
11. Meng L, Lin J, Chen L, Wang Z, Liu M, Liu Y, et al. Effectiveness and potential mechanisms of intralipid in treating unexplained recurrent spontaneous abortion. *Arch Gynecol Obstet* 2016; 294: 29-39.
12. Bepalova ON, Zhernakova TS, Selkov SA, Chepanov SV. Efficiency of intravenous therapy with Intralipid fat emulsion in patients with early reproductive loss. *J Obstet Womens Dis* 2022; 71: 21-8.
13. Peivandi S, Mortazavi L, Gordani N, Zamaniyan M, Asgarian-Omran H, Ajami A, et al. Effect of intralipid infusion on pregnancy outcome in infertile women with history of implantation failure: a single blind randomized clinical trial. *J Mazandaran Univ Med Sci* 2022; 32: 16-26.
14. Martini AE, Jasulaitis S, Fogg LE, Uhler ML, Hirshfeld-Cytron JE. Evaluating the utility of intralipid infusion to improve live birth rates in patients with recurrent pregnancy loss or recurrent implantation failure. *J Hum Reprod Sci* 2018; 11: 261-8.
15. Khairy M, Harb H, Eapen A, Melo P, Kazem R, Rajkhowa M, et al. The use of immunomodulation therapy in women with recurrent implantation failure undergoing assisted conception: a multicentre cohort study. *Am J Reprod Immunol* 2024; 91: e13819.
16. Kolanska K, Bendifallah S, Cohen J, Placais L, Selleret L, Johanne C, et al. Unexplained recurrent implantation failures: predictive factors of pregnancy and therapeutic management from a French multicentre study. *J Reprod Immunol* 2021; 145: 103313.
17. Placais L, Kolanska K, Kraiem YB, Cohen J, Suner L, Bornes M, et al. Intralipid therapy for unexplained recurrent miscarriage and implantation failure: case-series and literature review. *Eur J Obstet Gynecol Reprod Biol* 2020; 252: 100-4.
18. Mrosk P, Sandi-Monroy N, Gagsteiger F, Friedl TWP, Hancke

- K, Bundschu K. Endometrial scratching and intralipid treatment—no general recommendations. *Front Reprod Health* 2024; 6: 1505842.
19. Dogra Y, Singh N, Mathur S. Intralipid supplementation in women with unexplained recurrent implantation failure and elevated uterine natural killer cell levels—a randomized placebo controlled trial. *Hum Reprod* 2021; 29 Suppl 1: 307.
20. Gamaleldin I, Gomaa M, Shafik A, Akande V. Intralipid infusion does not improve live birth rates in women with unexplained recurrent implantation failure and may increase the risk of congenital malformations, a double-blinded randomised controlled trial. *BJOG* 2018; 125 Suppl 1: 31-2.
21. Rogenhofer N, Mahner S, von Hasselbach YL, Thaler CJ. Effects of intralipid infusions on anti-trophoblast antibody (ATAb)-activities in patients with recurrent pregnancy loss: an observational report. *Am J Reprod Immunol* 2022; 87: e13506.
22. Kumar P, Mahajan PMM. Estimation of peripheral NK cells (pNK) and role of intralipid and heparin in early recurrent pregnancy losses (RPL) – a pilot study. *J Assist Reprod Genet* 2016; 33: 1693-708.
23. Anya OC, Ajayi ER, Solanke HR, Ohanaka AI, Ubana KD, Ajayi E, et al. The use of intralipid infusions in the prevention of embryo implantation failure. *Cureus* 2024; 16: e53368.
24. Manning R, Irving L, Evbuomwan I. Add-on or take away: a retrospective review of managing women with recurrent implantation failure with intralipids during assisted reproductive cycles. *BJOG* 2019; 126 Suppl 1: 203.
25. Ehrlich R, Sarah G. A retrospective observational study of intralipid immunotherapy for women with reproductive failure. *J Assist Reprod Genet* 2016; 33: 1693-708.
26. Kaul V, Lolatgis N, Osianlis T, Sorby K, Vollenhoven B. The use of intralipid in IVF cycles in poor responders with repeated implantation failure who were positive for natural killer (NK) cells. *Hum Reprod* 2014; 29 Suppl 1: 128.
27. Holschbach V, Kuon R, Goeggel T, Daniel V, Weber M, Markert U, et al. Intravenous lipid administration in RPL and RIF patients with elevated uterine natural killer cells. *Hum Reprod* 2014; 29 Suppl 1: 140.
28. Marchand GJ, Masoud AT, Ulibarri H, Arroyo A, Coriell C, Goetz S, et al. Effect of a 20% intravenous fat emulsion therapy on pregnancy outcomes in women with RPL or RIF undergoing IVF/ICSI: a systematic review and meta-analysis. *J Clin Transl Res* 2023; 9: 236-45.
29. Zhou P, Wu H, Lin X, Wang S, Zhang S. The effect of intralipid on pregnancy outcomes in women with previous implantation failure in in vitro fertilization/intracytoplasmic sperm injection cycles: a systematic review and meta-analysis. *Eur J Obstet Gynecol Reprod Biol* 2020; 252: 187-92.
30. Rimmer MP, Black N, Keay S, Quenby S, Al Wattar BH. Intralipid infusion at time of embryo transfer in women with history of recurrent implantation failure: a systematic review and meta-analysis. *J Obstet Gynaecol Res* 2021; 47: 2149-56.
31. Han EJ, Lee HN, Kim MK, Lyu SW, Lee WS. Efficacy of intralipid administration to improve in vitro fertilization outcomes: a systematic review and meta-analysis. *Clin Exp Reprod Med* 2021; 48: 203-10.
32. Al-Zebeidi J, Agdi M, Lary S, Al-Obaid S, Salim G, Al-Jaroudi D. Effect of empiric intravenous intralipid therapy on pregnancy outcome in women with unexplained recurrent implantation failure undergoing intracytoplasmic sperm injection-embryo transfer cycle: a randomized controlled trial. *Gynecol Endocrinol* 2020; 36: 131-4.
33. Dakhly DM, Bayoumi YA, Sharkawy M, Gad Allah SH, Hassan MA, Gouda HM, et al. Intralipid supplementation in women with recurrent spontaneous abortion and elevated levels of natural killer cells. *Int J Gynaecol Obstet* 2016; 135: 324-7.

Beyond the funnel plot: editorial and reviewer perspectives on publication bias in systematic reviews and meta-analyses in clinical medicine

Hyun Kang

Department of Anesthesiology and Pain Medicine, Chung-Ang University College of Medicine, Seoul, Korea

Publication bias is a fundamental threat to the validity of systematic reviews and meta-analyses in clinical medicine. Yet current practice often reduces its assessment to the mechanical application of funnel plots, asymmetry tests, or single adjustment procedures, with limited attention to the underlying assumptions, alternative explanations, or implications for evidence certainty. This narrative methodological article reframes publication bias assessment as an interpretive and editorial responsibility rather than a purely technical problem. We examine what commonly used methods can and cannot reliably support. Detection tools function as nonspecific stress tests that identify deviations from simplified models; they do not diagnose selective publication but highlight situations in which the underlying assumptions require closer inspection. Adjustment approaches, including trim-and-fill, selection models, and regression-based methods, generate hypothetical estimates under unverifiable assumptions^{**}, and therefore provide^{**} sensitivity analyses rather than corrections that recover the true underlying effect. Divergence across adjustment methods is particularly informative, signaling inferential fragility rather than analytical failure. We identify five recurring misinterpretations encountered in peer review: equating asymmetry with proof of publication bias; privileging bias-adjusted estimates as inherently more credible; relying on a single adjustment method without examining assumption dependence; ignoring the plausibility of adjustment direction and magnitude; and overlooking implications for certainty of evidence. Editors and reviewers should prioritize transparency of assumptions, seriously consider alternative explanations, and calibrate conclusions proportionately. Viewing publication bias assessment as an interpretive responsibility rather than a methodological checklist promotes more disciplined inference and strengthens trust in clinical evidence synthesis.

Keywords: Evidence-based medicine; Meta-analysis as topic; Peer review, research; Publication bias; Research design; Systematic reviews as topic

Introduction

Publication bias has long been acknowledged as a fundamental threat to the validity of systematic reviews and meta-analyses in clinical research [1-4]. This refers to the selective processes by which studies with statistically significant or favorable results are more likely to be published, disseminated rapidly, or remain accessible, whereas studies with null or unfavorable findings are delayed, obscured, or

never published [1-6]. Closely related mechanisms—including time-lag bias, language bias, selective outcome reporting, and dissemination channel bias—operate along the same spectrum and collectively distort the evidentiary landscape available for synthesizing the evidence. The cumulative effect is a systematic inflation of pooled effect estimates and an overstatement of confidence in the resulting conclusions [1-6].

Although publication bias is widely recognized as an

Received: August 8, 2026; **Revised:** August 26, 2026; **Accepted:** September 8, 2026

Corresponding author: Hyun Kang

Email: roman00@naver.com

© 2026 Korean Society of Evidence-Based Medicine

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

important threat to the validity of systematic reviews and meta-analyses, it is often treated as a procedural rather than interpretative issue. During peer review, authors are commonly asked to show that publication bias has been assessed, usually by presenting a funnel plot or reporting the results of a statistical test. These analyses are frequently included as a routine requirement, with little consideration of their assumptions, limitations, alternative explanations for observed findings, or their implications for the certainty of the evidence [7-9]. Although numerous statistical methods have been developed to assess publication bias, ranging from classical funnel plot-based tests to more advanced approaches such as selection models and P curve analysis, there is still no consensus on how their results should be interpreted [1,10-13]. As a result, authors now have access to increasingly sophisticated analytical tools but often lack clear guidance on how to interpret and integrate these findings into balanced conclusions.

The Central Challenge: Interpretation, Not Methodology

Methodological literature on publication bias has largely focused on developing and comparing statistical tools, such as funnel plots, asymmetry tests, and bias-adjustment procedures [1,10-12]. While these methods are indispensable, their application alone does not ensure valid inference. This article contends that the central challenge of publication bias assessment is not methodological sufficiency, but interpretive discipline.

For example, a funnel plot may appear symmetric even when publication bias is present. Conversely, a statistically significant asymmetry test does not necessarily indicate publication bias, as asymmetry may result from factors unrelated to selective publication. Similarly, an adjusted effect estimate should not be considered more credible simply because it is closer to the null value [14,15]. Current evidence suggests that funnel plot asymmetry should be viewed as a nonspecific signal rather than direct evidence of publication bias [14,15]. Asymmetry may arise from genuine differences in treatment effects, variations in study quality, differences in outcome definitions, or differences in follow-up duration, even when no studies are missing. In other words, funnel plot asymmetry indicates that the observed data deviate from an idealized pattern, but it does not identify the reason for that deviation.

Scope and Organization

This article is organized into four sections. First, we review the available methods for detecting and adjusting for publication bias and clarify what conclusions these methods can support, distinguishing methods that provide sensitivity analyses from those that are often interpreted as diagnostic tests. Second, we discuss five common misinterpretations frequently encountered during peer review and explain why they occur. Third, we propose practical questions that editors and reviewers should consider when evaluating publication bias, with an emphasis on sound interpretation rather than simple methodological compliance. Finally, we discuss publication bias within certainty-of-evidence frameworks, particularly GRADE, and examine situations in which conventional methods provide limited information. Throughout this article, we emphasize that responsible assessment of publication bias requires transparent assumptions, careful consideration of alternative explanations, and conclusions that are consistent with the strength of the available evidence [16-18].

This article does not propose new statistical methods or advocate for any single analytical approach. Rather, it focuses on how existing methods should be interpreted, contextualized, and integrated into judgments about the certainty of evidence. The emphasis is placed on using detection and adjustment methods as interpretive stress tests and integrating their results into transparent, proportionate judgments about evidence certainty—moving beyond checklist compliance toward responsible evidence synthesis.

What Publication Bias Methods Actually Tell Us (and What They Do Not)

Publication bias methods are often presented as technical solutions to technical problems. From an editorial perspective, however, their principal value lies not in the numerical estimates they produce but in what they reveal about the limits of inference [19,20]. Therefore, understanding what these methods can—and, just as importantly, cannot—legitimately support is essential for fair, consistent, and proportionate peer review.

Detection tools: stress tests, not diagnostic tools

Detection tools, including funnel plots and small-study effect tests such as Egger's or Begg's tests, assess whether observed data are compatible with a simplified reference model—typically one assuming homogeneous effects and the absence of selective dissemination [21,22]. When funnel plot asymmetry or statistically significant small-study effects are observed,

the appropriate interpretation is that the reference model may be strained. Crucially, these findings do not identify the mechanism responsible for the deviation.

Consider a meta-analysis of surgical interventions, where smaller trials were conducted at specialized centers with more experienced surgeons, while larger pragmatic trials enrolled diverse centers. The observed asymmetry may reflect genuine effect modification by surgical expertise rather than selective publication. However, reviewers frequently interpret such patterns as confirmatory evidence of missing studies. This example illustrates a fundamental principle: funnel plot asymmetry may arise from true effect modification, methodological gradients, differences in outcome definitions, variations in follow-up duration, or systematic patterns of risk of bias—even when no studies are missing [13,14].

Accordingly, detection results should be interpreted as signals prompting further scrutiny rather than as evidence that publication bias has been demonstrated [21,22]. The appropriate question is not “Has publication bias been proven?” but rather “What might explain the observed pattern, and how much confidence can the evidence reasonably support given this uncertainty?”

Adjustment methods: sensitivity analyses, not corrections

Adjustment methods, including trim-and-fill, selection models, regression-based approaches such as PET-PEESE, and P value-based techniques, extend this logic by imposing explicit assumptions about how selective reporting might operate [23–25]. Each method answers a conditional question: If selection occurred according to a particular mechanism, how would the pooled estimate change? Therefore, the resulting adjusted estimates are hypothetical by construction. They provide information about sensitivity to assumptions but do not recover a true or unbiased underlying effect.

Divergent results across adjustment methods are common because each encodes a different and ultimately unverifiable narrative about the publication process [26]. To illustrate, consider how different adjustment methods respond to the same asymmetric funnel plot.

- Trim-and-fill might impute several small negative studies, shifting the pooled estimate toward the null.
- A selection model assuming P value-dependent publication might retain all observed studies but down-weight those with larger standard errors.
- Conversely, PET-PEESE extrapolates to zero precision using regression methods.

Each method encodes a distinct narrative of *how* selection operates. When these narratives yield divergent estimates,

the appropriate inference is not to average them or select the most conservative, but rather to acknowledge that conclusions depend critically on unverifiable assumptions about the publication process. This divergence signals inferential fragility rather than analytical failure.

From an editorial perspective, this divergence is deeply informative. Convergence across methods may suggest relative robustness across multiple assumptions, whereas disagreement signals inferential fragility and warrants restraint in interpretation. What matters is not whether an adjusted estimate is smaller, closer to the null, or statistically non-significant, but rather the transparency with which authors explain the drivers of change and moderate conclusions accordingly [1,27,28].

The distinction between detection, adjustment, and interpretation

In this sense, publication bias methods function most appropriately as interpretive stress tests rather than as corrective algorithms. They expose how sensitive conclusions are to plausible selection mechanisms, modeling choices, and analytical decisions [19,22]. Their primary contribution is not to identify a single preferred estimate but to delineate the range of conclusions that remain defensible under reasonable assumptions. Therefore, editors and reviewers should evaluate publication bias analyses not by their mere presence but by the transparency, coherence, and proportionality with which their implications are interpreted.

Table 1 contrasts these three components. Detection identifies when simplified models are strained, adjustment explores sensitivity to assumptions, and interpretation translates analytical findings into calibrated conclusions. Conflating these stages—for example, treating detection results as proof of publication bias or adjustment outputs as corrections—represents the most consequential category error in publication bias assessment.

Together, these components demonstrate that publication bias assessment is not a linear progression toward a corrected estimate, but rather a process by which analytical findings constrain the range of defensible inferences.

Common Misinterpretations Encountered in Peer Review

Although publication bias is widely acknowledged as a threat to evidence synthesis, its assessment is frequently misunderstood during the peer review process. These misunderstandings rarely arise from a lack of statistical knowledge. Instead, they reflect a more fundamental problem: the conflation

Table 1. Detection, Adjustment, and Interpretation in the Editorial Evaluation of Publication Bias

Dimension	Detection Methods	Adjustment Methods	Interpretation (Editorial Task)
Primary purpose	Identify deviations from a simplified reference model	Explore how pooled estimates change under assumed selection mechanisms	Judge how much confidence the evidence can reasonably support
Typical examples	Funnel plots; Egger's test; Begg's test	Trim-and-fill; selection models; PET-PEESE; P value-based methods	Editorial synthesis of bias analyses, heterogeneity, and study quality
Core question addressed	Do the observed data fit a model with no selection or homogeneous effects?	If selection operates in a specific way, how would the estimate change?	How should conclusions be calibrated, given assumption dependence and inferential fragility?
Nature of output	Graphical patterns or test statistics	Hypothetical bias-adjusted effect estimates	Judgments about certainty of evidence and strength of claims
Key assumptions	Homogeneous effects; absence of selective dissemination	Explicit, method-specific selection mechanisms	Transparency, plausibility, and proportionality of inference
What the results can tell us	That the simple reference model may be violated	That conclusions are sensitive (or robust) to particular assumptions	Whether confidence should be downgraded or conclusions moderated
What the results cannot tell us	The cause of asymmetry or deviation	The true underlying effect size	A single 'corrected' or definitive estimate
Common misinterpretation	Asymmetry interpreted as proof of publication bias	Adjusted estimate treated as closer to the truth	Confidence maintained despite demonstrated fragility
Appropriate editorial use	Prompt further scrutiny and exploration of alternatives	Function as sensitivity analyses, not replacements	Integrate findings into certainty-of-evidence judgments (e.g., GRADE)
Risk if overemphasized	Overdiagnosis of publication bias	False sense of correction or precision	Overconfident conclusions despite uncertainty

PET-PEESE: precision-effect test and precision-effect estimate with standard error, GRADE: Grading of Recommendations, Assessment, Development, and Evaluations.

of detection, adjustment, and interpretation as if they were interchangeable steps toward a single, corrected estimate. Predictable errors occur when publication bias methods are treated as diagnostic or corrective tools rather than as components of a broader interpretive process. Five recurring patterns merit particular attention.

Misinterpretation 1: equating asymmetry with proof of publication Bias

Error: One of the most common errors is to equate funnel plot asymmetry or a statistically significant small-study effect test with the presence of publication bias. In peer review, this often appears as statements such as, “Egger’s test is significant; therefore, publication bias exists.” Such reasoning treats detection results as causal diagnoses, overlooking the fact that asymmetry merely indicates a deviation from a simplified reference model assuming homogeneous effects and no selection; it does not explain why that deviation occurred.

Why This Matters: Interpreting asymmetry as proof of selective non-publication collapses multiple plausible explanations into a single unsupported conclusion [1,12]. An evidence base may be fully published and yet still display systematic asymmetry. This misinterpretation leads to an unwarranted downgrading of evidence certainty based on

ambiguous signals.

Editorial Response: When asymmetry is detected, editors should ask: “What alternative explanations have the authors considered? Have they examined heterogeneity, methodological quality gradients, and differences in outcome definitions? Are conclusions appropriately cautious given this uncertainty?”

Misinterpretation 2: treating bias-adjusted estimates as inherently more credible

Error: A second critical misinterpretation treats bias-adjusted estimates as inherently more credible than unadjusted results, which is a logical error that misunderstands the hypothetical nature of adjustment procedures. This is often reflected in reviewer requests to present trim-and-fill or similar adjusted results as primary estimates. Such requests disregard the fact that adjustment methods generate estimates under specific unverifiable assumptions about selection.

Why This Matters: An effect estimate that is smaller or closer to the null is not inherently less biased; it may simply reflect the mathematical structure of the chosen method rather than any improvement in evidential validity [1,19,20,23]. Privileging attenuated estimates risks substituting one untested assumption for another rather than

improving inference. When an adjusted estimate is smaller, reviewers may feel reassured without scrutinizing whether the implied adjustment is substantively credible.

Editorial Response: Editors should ask authors to present unadjusted and adjusted estimates side by side and to state explicitly which assumption drives the difference between them. An adjusted estimate should be reported as one scenario among several, not as the primary result. Where the adjustment is presented as the headline finding, editors should request that the framing be reversed.

Misinterpretation 3: treating a single adjustment method as definitive

Error: Peer reviewers may focus on whether a publication bias method was applied rather than on how the results behave across multiple plausible approaches. When a single adjusted estimate is presented without a comparison to alternatives, or when divergent results across methods are not discussed, a sensitivity analysis is implicitly elevated to a corrective conclusion. This practice obscures assumption dependence and exaggerates certainty.

Why This Matters: This error is particularly consequential in the presence of substantial between-study heterogeneity, where different adjustment methods encode fundamentally different narratives about selection mechanisms [19,28]. What should function as an exploration of fragility instead becomes a misplaced endpoint, creating false confidence in the single estimate.

Editorial Response: Editors should request that authors present results across multiple adjustment approaches and explicitly discuss how the estimates differ. Stability across diverse methods may support conditional confidence, while divergence signals inferential fragility, which should be reflected in more cautious language and lower certainty-of-evidence ratings.

Misinterpretation 4: ignoring the plausibility of adjustment direction and magnitude

Error: Less frequently articulated but equally consequential is the failure to examine whether the direction and magnitude of an adjustment are substantively plausible. Attenuation toward the null is often accepted as reassuring without scrutiny of whether the implied missing or down-weighted studies resemble realistic trials.

Why This Matters: Consider a meta-analysis of a novel pharmacological intervention in which all included studies were industry-sponsored and showed large effects. Trim-and-fill analysis may impute several small negative studies to achieve symmetry. However, if no such negative studies exist

in the evidence base despite extensive searching, the imputed studies may be implausibly pessimistic. Adjustments that shift influence toward studies unlike any observed in the literature should prompt concern rather than confidence [19,23].

Editorial Response: When adjustment methods are employed, editors should verify that (1) the implied selection mechanism reflects plausible publication dynamics in the field, (2) hypothetical imputed studies resemble realistic trials, and (3) the adjustment does not systematically disadvantage methodologically rigorous but smaller studies.

Misinterpretation 5: overlooking implications for certainty of evidence

Error: Finally, publication bias analyses are often discussed in isolation without explicit consequences for interpretation. Even when adjusted estimates vary widely or demonstrate strong dependence on modeling assumptions, manuscripts may retain confident causal language and strong clinical recommendations. In such cases, the central issue is not whether publication bias was “detected,” but whether its potential impact has been appropriately translated into judgments about the certainty of evidence and the strength of conclusions [20,28–30].

Why This Matters: This disconnection between analytical findings and interpretive restraint represents a fundamental breakdown in the evidence synthesis discipline. When the analysis reveals inferential fragility, the conclusions must correspondingly weaken. Maintaining strong recommendations despite demonstrated sensitivity to unverifiable assumptions violates the principle of proportionality [29,32].

Editorial Response: Editors should assess whether strong assumption dependence, wide sensitivity ranges, or instability across models are reflected in more cautious language, explicit downgrading of certainty of evidence per GRADE framework, and moderated recommendations. When analytical fragility is documented but conclusions remain confident, an unresolved disconnect persists between the analysis and interpretation.

Summary: These five recurring misinterpretations highlight the need for an explicit editorial framework that distinguishes detection from adjustment and both from interpretation. Shifting attention away from method selection and toward inferential judgment, transparency, and proportional interpretation is essential if publication bias analyses are to meaningfully inform evidence appraisal rather than merely satisfying procedural expectations.

Editorial Questions That Matter More Than the Method Used

From an editorial perspective, the central task in evaluating publication bias analyses is not to verify whether specific statistical procedures have been applied but to determine whether the analysis demonstrates sound inferential reasoning. This requires shifting attention away from method labels and toward the logic that connects the assumptions, analytical results, and resulting conclusions. In practice, these questions function as safeguards against the category errors outlined above, particularly the conflation of detection, adjustment, and interpretation [19,31].

Five Essential Editorial Questions

What assumptions does the analysis rely on, and are they explicitly stated?

All publication bias methods encode assumptions regarding how studies are selected, reported, or disseminated. Therefore, editorial evaluation should begin by asking whether these assumptions are clearly articulated and whether they plausibly reflect the clinical and publication contexts of the evidence base. Analyses that present bias-adjusted estimates without explaining the implied selection mechanism provide little interpretive grounding regardless of statistical sophistication. Without explicit assumptions, neither the detection nor the adjustment results can be meaningfully interpreted, and any downstream conclusions remain unsupported [4,20,31].

Are alternative explanations for the observed patterns seriously considered?

When funnel plot asymmetry or small-study effects are reported, authors should address explanations beyond selective non-publication. These include true effect modification, gradients in methodological quality, differences in outcome definitions, variations in follow-up duration, and analytical flexibility. Editorial judgment should favor analyses that situate publication bias within this broader explanatory landscape rather than treating it as a default diagnosis or a standalone issue. Failure to consider alternatives transforms detection signals into causal claims, reinforcing the misinterpretations commonly encountered in peer reviews [1,11,19,32].

How stable are the conclusions across reasonable analytical choices?

A robust publication bias assessment makes sensitivity visi-

ble. Editors should look for transparent comparisons across multiple plausible adjustment methods or model specifications, including scenarios in which the conclusions weaken, attenuate, or change direction. Stability across approaches may support conditional confidence, whereas divergence indicates inferential fragility. Crucially, both outcomes are informative only if they are reported and interpreted and not suppressed in favor of a single adjusted estimate [20,28,32].

Have the direction and magnitude of the adjustment been critically examined?

Noting that an adjusted estimate is smaller or closer to the null hypothesis is insufficient. Editorial scrutiny should consider whether implied adjustments are substantively credible. Do the hypothetical missing or down-weighted studies resemble plausible trials in the field? Does the adjustment disproportionately privilege larger but lower-quality studies over smaller, methodologically rigorous ones? Unexamined attenuation toward the null risks substituting one untested assumption for another rather than improving inference [20,28,32].

Are the consequences for the certainty of evidence clearly articulated?

Most importantly, publication bias analysis should have visible consequences for interpretation. Editors should assess whether strong assumption dependence, wide sensitivity ranges, or instability across models are reflected in more cautious language, explicit downgrading of the certainty of evidence, or moderated recommendations. When analytical fragility is documented but conclusions remain confident, an unresolved disconnect persists between analysis and interpretation. At this stage, publication bias assessment ceases to be a methodological issue and becomes a question of evidential responsibility [30,33].

These questions emphasize that publication bias assessment is fundamentally an exercise in interpretation rather than algorithm selection. Editorial evaluation should therefore prioritize transparency, restraint, and coherence of reasoning over mechanical compliance with the methodological checklists. Viewed through this lens, publication bias analysis directly informs judgments about the certainty of evidence rather than functioning as a stand-alone diagnostic step or a means of effect size correction.

Publication Bias as a Certainty-of-Evidence Issue

Within evidence-based medicine, publication bias is best

understood not as a statistical inconvenience but as a fundamental threat to the certainty of evidence. Frameworks such as GRADE explicitly recognize publication bias as a reason to downgrade confidence in effect estimates, reflecting the insight that selective dissemination distorts the evidentiary base itself rather than merely shifting numerical results [18,33]. From this perspective, the goal of publication bias assessment is not to repair an estimate but to determine how cautiously that estimate should be interpreted.

Publication bias as evidence fragility

This reframing has direct implications for editorial decision making. Inferential fragility is indicated when bias-adjusted estimates vary widely across plausible assumptions, results are highly sensitive to unverifiable selection mechanisms, or conclusions hinge on a small number of influential studies. In GRADE terms, such fragility justifies downgrading the certainty of evidence, even when statistical significance persists under the selected models. The apparent numerical robustness within a single analytical framework does not override the instability across reasonable alternatives [18,33].

Equally important, the absence of statistically significant evidence of publication bias should not be interpreted as reassurance. Funnel plot symmetry, non-significant asymmetry tests, or convergence of a single adjustment method merely indicate that selective processes were not detected under the assumptions applied. They do not establish that publication bias is absent. Treating “no evidence of bias” as evidence of no bias conflates detection with interpretation, which is a mistake particularly consequential in settings characterized by substantial heterogeneity, flexible outcome definitions, or complex reporting practices, where conventional detection tools are known to be weakly informative [7,12,18].

When evidence can retain higher certainty

Conversely, when publication bias analyses demonstrate relative stability across multiple methods and assumptions, especially in evidence bases with prespecified outcomes, low heterogeneity, and external constraints such as trial registration, editors may reasonably judge the risk of publication bias to be lower. However, even under these circumstances, the appropriate inference is conditional confidence rather than confirmation of a true effect. Certainty may be retained, but it remains explicitly contingent on the assumptions examined [19,21].

Integration with GRADE framework

Viewed through a certainty-of-evidence lens, publication bias analysis functions as the final interpretive step linking

statistical modeling to editorial responsibility. Its primary contribution lies in translating analytical sensitivity into calibrated language, moderated recommendations, and explicit acknowledgment of uncertainties. At this stage, publication bias assessment ceases to be a methodological exercise and becomes an ethical one, requiring that claims be proportionate to the robustness of the evidence supporting them [18,33].

In this sense, publication bias analysis does not tell us which estimate to trust; it tells us how much trust the evidence can reasonably bear.

Special Contexts: When Editors and Reviewers Should Watch Carefully

Although publication bias is a general concern in all forms of evidence synthesis, certain research contexts pose a heightened risk of misinterpretation. In these settings, conventional detection or adjustment methods may appear reassuring while obscuring deeper forms of selective dissemination and analytical flexibility. For editors and reviewers, the challenge is not the absence of methods but the presence of apparently coherent results in contexts where inferential fragility is structurally amplified [15,34] (Table 2).

Umbrella reviews and higher-order syntheses

Umbrella reviews synthesize evidence from multiple meta-analyses and therefore operate at an increased distance from the primary study selection. This layered structure magnifies the consequences of selective dissemination, rather than neutralizing them. Publication and reporting biases at the primary study level may propagate through individual meta-analyses and be further reinforced when umbrella reviews preferentially cite statistically significant or highly cited syntheses [35,36].

In addition, the overlap of primary studies across meta-analyses creates an illusion of independent replication, as the same influential trials—often those with large or favorable effects—are repeatedly reintroduced. In such settings, funnel plot-based assessments conducted at the meta-analysis level may appear well-behaved even when the underlying evidence base is highly selective, requiring editorial scrutiny that extends beyond per-meta-analysis diagnostics [35,36].

Flexible outcomes, time points, and analytical choices

Publication bias is commonly conceptualized as the non-publication of entire studies. However, in many clinical studies, the dominant selective process operates within studies rather than between them. Selective reporting of outcomes, time points, subgroups, or analytical models can generate

Table 2. Domain and Key Questions Editors and Reviewers Should Ask

Domain	Key Questions Editors and Reviewers Should Ask
Conceptual framing	Do the authors clearly explain what publication bias methods can and cannot tell us? Are adjustment methods framed as sensitivity analyses rather than as causal corrections?
Detection	Are funnel plots or small-study effect tests interpreted as stress tests rather than proof of publication bias? Are alternative explanations, such as heterogeneity or risk-of-bias gradients, discussed?
Choice of methods	Is the choice of publication bias methods justified in relation to plausible selection mechanisms in the field, rather than being applied mechanically?
Multiplicity of models	Were multiple adjustment approaches presented when publication bias was suspected? If the results differ, is this divergence highlighted and interpreted?
Transparency of assumptions	Do the authors clearly describe the assumptions encoded by each method (e.g., symmetry, P value dependence, precision–effect relationships)?
Integration with heterogeneity	Are publication bias analyses interpreted alongside heterogeneity, risk of bias assessments, and information size, rather than in isolation?
Certainty of evidence	Are the findings from publication bias analyses explicitly linked to judgments about the certainty of the evidence and the strength of the conclusions?
Scope and limits	Do the authors acknowledge when publication bias methods are weakly informative, such as when few studies or substantial heterogeneity are present
Reporting balance	Are both unadjusted and adjusted results presented transparently, without privileging a single 'corrected' estimate?
Conclusion discipline	Do the conclusions appropriately reflect evidential fragility when the results are sensitive to assumptions, or do they retain strong causal language despite uncertainty?

robust pooled effects, even when all studies are published [13,36].

Conventional publication bias methods are poorly equipped to detect such within-study selection, and funnel plot symmetry or stable-adjusted estimates may therefore provide false reassurance. Editors should be particularly alert when outcome definitions are heterogeneous, follow-up windows are flexible, or analytical decisions are weakly pre-specified, as these features expand the space for selective reporting [15,37,38].

Preprints, living reviews, and rapidly evolving literatures

The increasing use of preprints and living systematic reviews has altered the dynamics of the selective dissemination of research. The inclusion of preprints may mitigate traditional publication bias by capturing studies that have not yet been filtered by journal publication. Simultaneously, rapid dissemination environments may preferentially amplify novel, positive, or attention-grabbing findings, introducing new forms of selection.

Editorial evaluation should therefore focus not on whether preprints are included but on how their inclusion affects interpretability. Analyses that explicitly examine the influence of preprints on pooled estimates provide more informative evidence than their unexamined inclusion or exclusion [31,39].

Automation and AI-assisted evidence synthesis

Automation and artificial intelligence–assisted tools are increasingly used for literature screening, prioritization, and synthesis. While these tools improve efficiency, they may also inherit and amplify existing publication and citation biases. Algorithms trained predominantly on published or highly cited studies risk reinforcing selective visibility, rather than correcting it.

Transparency is essential from an editorial perspective. Editors and reviewers should assess whether the automated processes were audited, whether the screening and prioritization decisions were reproducible, and whether human judgment was retained at the interpretive stage, particularly in the publication bias assessment [40,41].

Small evidence bases and influential trials

Finally, publication bias methods are especially fragile when the number of included studies is small or when a few trials dominate the evidence. In such circumstances, both detection and adjustment procedures can behave erratically, with single influential studies driving apparent asymmetry or large shifts in the adjusted estimates. Analytical sophistication cannot compensate for the lack of evidence.

Therefore, editorial judgment should favor the explicit acknowledgment of uncertainty over reliance on any single bias-adjustment procedure [1,15]. Transparency about these limitations is more informative than false confidence in methods applied to inadequate evidence bases.

These contexts underscore that publication bias assess-

ment cannot be interpreted independently of the evidence base's structure and dynamics. In complex, flexible, or rapidly evolving literatures, restraint, transparency, and attention to the certainty of evidence are often more informative than elaborate modeling. For editors and reviewers, recognizing when methods are weakly informative is as important as their correctness.

An editorial checklist for publication bias assessment

Rather than prescribing specific statistical methods, the following checklist emphasizes interpretive quality, transparency and inferential discipline. Each domain reflects a question that editors and reviewers should consider when assessing whether publication bias analyses meaningfully inform the interpretation.

This checklist is intentionally brief and interpretive in nature. Its purpose is not to standardize which methods must be used but to ensure that how methods are used aligns with what they can legitimately support.

Conclusions: Publication Bias as an Interpretive Responsibility

Publication bias methods do not correct bias in a causal sense. When used well, they reveal how dependent clinical conclusions are on unverifiable assumptions about selection, reporting, and study visibility. Divergent adjusted estimates, wide sensitivity ranges, and strong dependence on modeling choices are not analytical failures; they are signals of evidential fragility. Recognizing this distinction transforms publication bias assessment from a technical problem into an interpretive responsibility.

For authors, responsible use of publication bias methods means resisting the temptation to present a single “corrected” estimate and instead documenting how conclusions change across plausible scenarios. For reviewers, this means evaluating coherence and transparency rather than technical box-checking. For editors, this means recognizing that the most important question is not whether publication bias was tested but whether uncertainty was honestly carried through to interpretation and claims.

Viewed through a certainty-of-evidence framework such as GRADE, the primary contribution of publication bias analysis is to inform the appropriate downgrading of confidence. When conclusions hinge on strong, unverifiable assumptions, certainty should be reduced accordingly, even if numerical estimates appear precise. In this sense, publication

bias analysis serves as a bridge between statistical modeling and responsible interpretation of the results.

Treating publication bias as an interpretive responsibility rather than a statistical obstacle helps ensure that meta-analyses support clinical recommendations that are proportional to the strength of the underlying evidence. This shift from correction to calibration is essential for preserving trust in evidence synthesis in an era of increasing analytical complexity and automation. By prioritizing transparency, coherence, and proportional inference over mechanical compliance with analytical checklists, editors and reviewers can ensure that publication bias assessment meaningfully informs evidence appraisal and strengthens, rather than undermines, confidence in clinical evidence synthesis.

Conflict of Interest

Hyun Kang has been the Editor-in-Chief of the Journal of Evidence-Based Practice since 2025; however, he was not involved in the peer reviewer selection, evaluation, or decision process of this article. No other potential conflicts of interest relevant to this article were reported.

Funding

This research was funded by the Basic Science Research Program through the National Research Foundation (NRF) of Korea, funded by the Ministry of Education, Science, and Technology (NRF-2022R1F1A1074934).

Data Availability Statement

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Ethics Approval and Consent to Participate

Not applicable.

Authors' Contributions

Conceptualization: HK. Funding acquisition: HK. Methodology: HK. Writing – original draft: HK. Writing – review & editing: HK.

Acknowledgments

None.

ORCID

Hyun Kang, <https://orcid.org/0000-0003-2844-5880>

References

1. Debray TPA, Moons KGM, Riley RD. Detecting small-study effects and funnel plot asymmetry in meta-analysis of survival data: A comparison of new and existing tests. *Res Synth Methods* 2018; 9: 41-50.
2. Sterne JA, Egger M, Smith GD. Systematic reviews in health care: investigating and dealing with publication and other biases in meta-analysis. *BMJ* 2001; 323: 101-5.
3. Dwan K, Altman DG, Arnaiz JA, Bloom J, Chan AW, Cronin E, et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS One* 2008; 3: e3081.
4. Hopewell S, Loudon K, Clarke MJ, Oxman AD, Dickersin K. Publication bias in clinical trials due to statistical significance or direction of trial results. *Cochrane Database Syst Rev* 2009; 1: MR000006.
5. Marks-Anglin A, Chen Y. A historical review of publication bias. *Res Synth Methods* 2020; 11: 725-42.
6. Cheema HA, Shahid A, Ehsan M, Ayyan M. The misuse of funnel plots in meta-analyses of proportions: are they really useful? *Clin Kidney J* 2022; 15: 1209-10.
7. Afonso J, Ramirez-Campillo R, Clemente FM, Buttner FC, Andrade R. The perils of misinterpreting and misusing "publication bias" in meta-analyses: an education review on funnel plot-based methods. *Sports Med* 2024; 54: 257-69.
8. Hamilton DG, Fraser H, Hoekstra R, Fidler F. Journal policies and editor's opinions on peer review. *eLife* 2020; 9: e62529.
9. Page MJ, Higgins JP, Sterne JA. Assessing risk of bias due to missing results in a synthesis. editors. *Cochrane Handbook for Systematic Reviews of Interventions*. Version 6.5. In: Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, , London, Cochrane. 2024.
10. Sterne JA, Egger M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. *J Clin Epidemiol* 2001; 54: 1046-55.
11. Zwetsloot PP, Van Der Naald M, Sena ES, Howells DW, Int'Hout J, De Groot JA, et al. Standardized mean differences cause funnel plot distortion in publication bias assessments. *Elife* 2017; 6: e24260.
12. Doleman B, Freeman SC, Lund JN, Williams JP, Sutton AJ. Funnel plots may show asymmetry in the absence of publication bias with continuous outcomes dependent on baseline risk: presentation of a new publication bias test. *Res Synth Methods* 2020; 11: 522-34.
13. Kang H. Correcting what cannot be corrected: rethinking publication bias analysis methods in clinical meta-analyses. *Korean J Anesthesiol* 2026; 79: 271-90.
14. Muradchanian J, Hoekstra R, Kiers H, van Ravenzwaaij D. The role of results in deciding to publish: A direct comparison across authors, reviewers, and editors based on an online survey. *PLoS One* 2023; 18: e0292279.
15. Mathur MB, VanderWeele TJ. Estimating publication bias in meta-analyses of peer-reviewed studies: A meta-meta-analysis across disciplines and journal tiers. *Res Synth Methods* 2021; 12: 176-91.
16. Schünemann HJ, Neumann I, Hultcrantz M, Brignardello-Pettersen R, Zeng L, Murad MH, et al. GRADE guidance 35: update on rating imprecision for assessing contextualized certainty of evidence and making decisions. *J Clin Epidemiol* 2022; 150: 225-42.
17. Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. *Cochrane Handbook for Systematic Reviews of Interventions*. Version 6.5. London, Cochrane. 2024.
18. Prasad M. Introduction to the GRADE tool for rating certainty in evidence and recommendations. *Clin Epidemiol Glob Health* 2024; 25: 101484.
19. Mavridis D, Salanti G. How to assess publication bias: funnel plot, trim-and-fill method and selection models. *Evid Based Ment Health* 2014; 17: 30.
20. Lin L, Chu H. Quantifying publication bias in meta-analysis. *Biometrics* 2018; 74: 785-94.
21. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997; 315: 629-34.
22. Möller AP, Jennions MD. Testing and adjusting for publication bias. *Trends Ecol Evol* 2001; 16: 580-6.
23. Shi L, Lin L. The trim-and-fill method for publication bias: practical guidelines and recommendations based on a large database of meta-analyses. *Medicine (Baltimore)* 2019; 98: e15987.
24. Maier M, VanderWeele TJ, Mathur MB. Using selection models to assess sensitivity to publication bias: A tutorial and call for more routine use. *Campbell Syst Rev* 2022; 18: e1256.
25. Alinaghi N, Reed WR. Meta-analysis and publication bias: how well does the FAT-PET-PEESE procedure work? *Res Synth Methods* 2018; 9: 285-311.
26. Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Comparison of two methods to detect publication bias in meta-analysis. *JAMA* 2006; 295: 676-80.
27. Bartoš F, Maier M, Quintana DS, Wagenmakers E-J. Adjusting for publication bias in JASP and R: selection models, PET-PEESE, and robust Bayesian meta-analysis. *Adv Methods Pract Psychol Sci* 2022; 5: 25152459221109259.

28. Ahmed I, Sutton AJ, Riley RD. Assessment of publication bias, selection bias, and unavailable data in meta-analyses using individual participant data: a database survey. *BMJ* 2012; 344: d7762.
29. Mathur MB. Assessing robustness to worst case publication bias using a simple subset meta-analysis. *BMJ* 2024; 384: e076851.
30. Al Duhailib Z, Granholm A, Alhazzani W, Oczkowski S, Belsey-Cote E, Møller MH. GRADE pearls and pitfalls-part 1: systematic reviews and meta-analyses. *Acta Anaesthesiol Scand* 2024; 68: 584-92.
31. Page MJ, Sterne JAC, Higgins JPT, Egger M. Investigating and dealing with publication bias and other reporting biases in meta-analyses of health research: A review. *Res Synth Methods* 2021; 12: 248-59.
32. Formann AK. Estimating the proportion of studies missing for meta-analysis due to publication bias. *Contemp Clin Trials* 2008; 29: 732-9.
33. Daou JP, Riera R, Pacheco RL. Reasons for downgrading the certainty of evidence for indirectness in synthesis of surgical procedures for patients with fractures: a meta-research analysis. *J Eval Clin Pract* 2025; 31: e70091.
34. Ropovik I, Adamkovic M, Greger D. Neglect of publication bias compromises meta-analyses of educational research. *PLoS One* 2021; 16: e0252415.
35. Fusar-Poli P, Radua J. Ten simple rules for conducting umbrella reviews. *Evid Based Ment Health* 2018; 21: 95-100.
36. Choi GJ, Kang H. The umbrella review: a useful strategy in the rain of evidence. *Korean J Pain* 2022; 35: 127-8.
37. Song F, Parekh S, Hooper L, Loke YK, Ryder J, Sutton AJ, et al. Dissemination and publication of research findings: an updated review of related biases. *Health Technol Assess* 2010; 14: iii, ix-xi, 1-193.
38. Lindsley K, Fusco N, Li T, Scholten R, Hooft L. Clinical trial registration was associated with lower risk of bias compared with non-registered trials among trials included in systematic reviews. *J Clin Epidemiol* 2022; 145: 164-73.
39. Kang H. Beyond the paywall: the role of preprints in overcoming publication bias. *J Evid-Based Pract* 2025; 1: 7-11.
40. Moens M, Nagels G, Wake N, Goudman L. Artificial intelligence as team member versus manual screening to conduct systematic reviews in medical sciences. *iScience* 2025; 28: 113559.
41. Millard LA, Flach PA, Higgins JP. Machine learning to assist risk-of-bias assessments in systematic reviews. *Int J Epidemiol* 2016; 45: 266-77.

Evolution of evidence-based medicine: toward digital evidence ecosystems and artificial intelligence

Soo Young Kim

Department of Family Medicine, Kangdong Sacred Heart Hospital, Hallym University College of Medicine, Seoul, Korea

Evidence-based medicine (EBM) has transformed clinical decision-making by integrating the best available research evidence with clinical expertise and patients' values and preferences. Over the past three decades, EBM has evolved from the critical appraisal of individual studies into a broader framework encompassing evidence synthesis, clinical practice guidelines, assessment of evidence certainty, research transparency, and shared decision-making. Despite these advances, contemporary EBM faces important challenges, including the rapidly increasing volume of research, delays in evidence synthesis and implementation, limited applicability of randomized controlled trials to heterogeneous real-world populations, and difficulties in individualizing population-level evidence. Emerging approaches—including real-world evidence, living evidence, learning health systems, precision medicine, and artificial intelligence (AI)—offer opportunities to address these limitations. Together, these approaches may enable a transition from static to continuously updated evidence, from population-average to more personalized evidence, and from a linear evidence pipeline to a learning evidence ecosystem in which clinical practice both uses and generates evidence. AI may further accelerate evidence retrieval, synthesis, updating, and individualized decision support, while introducing challenges related to reliability, bias, transparency, reproducibility, and accountability. Next-generation EBM should therefore be conceptualized not as a replacement for traditional EBM but as its evolution into a digitally connected, continuously learning evidence ecosystem. In the AI era, the foundational principles of EBM—source verification, critical appraisal, uncertainty assessment, integration of patient preferences, and accountable human judgment—will become increasingly important.

Keywords: Artificial Intelligence; Evidence-Based Medicine; Learning Health System; Precision Medicine; Real-World Evidence

Introduction

Evidence-based medicine (EBM) has transformed clinical decision-making over the past three decades. EBM emerged in response to growing concerns about variation in clinical practice, the underuse of effective interventions, the overuse of unnecessary care, and the difficulty of translating rapidly expanding medical research into clinical decisions [1,2].

EBM introduced a systematic approach to clinical decision-making by integrating the best available research evidence with clinical expertise and patients' values and preferences [3]. It promoted the formulation of answerable clinical

questions, systematic identification and critical appraisal of relevant evidence, and explicit application of evidence to patient care. Over time, EBM expanded beyond the appraisal of individual studies to include systematic reviews, clinical practice guidelines, structured assessment of the certainty of evidence, and shared decision-making, thereby strengthening the transparency and consistency of clinical decisions [4,5].

Despite these advances, important challenges remain. The volume and pace of research now exceed clinicians' capacity to identify and evaluate relevant evidence, while systematic reviews and guidelines may become outdated as new

Received: August 11, 2026; **Revised:** September 1, 2026; **Accepted:** September 11, 2026

Corresponding author: Soo Young Kim

Email: hallymfm@gmail.com

© 2026 Korean Society of Evidence-Based Medicine

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

evidence emerges [6]. Moreover, average treatment effects derived from randomized controlled trials (RCTs) may not adequately inform decisions for individual patients, particularly older adults with multimorbidity, polypharmacy, or characteristics that are underrepresented in clinical trials.

These challenges have stimulated growing interest in real-world evidence (RWE), living evidence, learning health systems (LHSs), precision medicine, and artificial intelligence (AI) [7]. Rather than replacing EBM, these approaches may extend its principles by enabling evidence to be generated, synthesized, updated, and applied more continuously and efficiently. This review examines the evolution and current limitations of EBM and explores how these emerging approaches may contribute to a next-generation digital evidence ecosystem.

Emergence and Core Principles of Evidence-Based Medicine

Emergence of EBM

The development of EBM was shaped by the work of Archie Cochrane, David Sackett, Gordon Guyatt, and others who advocated the systematic use of empirical evidence in clinical decision-making. The term evidence-based medicine became widely used in the early 1990s through the work of Guyatt and colleagues, who described EBM as a new paradigm for medical practice and education [1,2]. The subsequent development of the Cochrane Collaboration, systematic reviews, evidence-based clinical practice guidelines, and the GRADE approach extended EBM from the critical appraisal of individual studies to systematic evidence synthesis and explicit recommendations [2,4,5].

Three Components and Five Steps of EBM

EBM integrates the best available research evidence with clinical expertise and patients' values and preferences [3]. Clinical decisions therefore depend not only on research findings but also on individual patient circumstances, expected benefits and harms, feasibility, and patient priorities [3,8].

EBM is commonly practiced through five steps: Ask, Acquire, Appraise, Apply, and Assess. These steps provide a practical framework for translating clinical uncertainty into evidence-informed decisions and have also become a foundation for EBP education [8]. The Sicily Statement emphasized that EBP requires not only knowledge but also the skills and attitudes necessary to formulate questions, search for and critically appraise evidence, apply it in practice, and evaluate performance [8]. Albarqouni and colleagues

subsequently developed a consensus-based set of core EBP competencies encompassing question formulation, evidence searching and appraisal, interpretation of effects and uncertainty, application of evidence, and integration of patient values and preferences [9]. EBM can therefore be understood not only as an approach to using research evidence but also as a set of competencies for rational decision-making under uncertainty [8,9].

Expansion of EBM

EBM has expanded from individual clinicians' appraisal of research toward a broader system encompassing evidence generation, synthesis, recommendations, implementation, and evaluation, and from medicine toward evidence-based practice across health professions [2,8].

Systematic reviews became central to evidence synthesis, while GRADE provided a structured framework for assessing the certainty of evidence and distinguishing it from the strength of recommendations [4,10]. The subsequent development of GRADE Evidence-to-Decision frameworks further emphasized that recommendations should consider not only the certainty of evidence but also the balance of benefits and harms, patients' values and preferences, resource use, equity, acceptability, feasibility, and context [11].

EBM consequently evolved from the appraisal of individual studies into a broader system in which evidence is systematically synthesized, translated into recommendations, and incorporated into clinical and health care decisions [2,5,11].

Transparency, Reproducibility, and Patient-Centeredness

Research registration, reporting guidelines, data sharing, open science, and shared decision-making have further expanded EBM by improving the transparency, reproducibility, and patient-centeredness of research and clinical decisions. Reporting guidelines such as CONSORT, STROBE, STARD, and PRISMA promote more complete and transparent reporting, thereby facilitating critical appraisal and evidence synthesis [12]. Trial registration and protocol availability can help identify selective reporting and undisclosed changes in study outcomes or methods [13].

The FAIR principles emphasize that scientific data should be Findable, Accessible, Interoperable, and Reusable [14], an increasingly important requirement for digital and AI-based evidence systems. Shared decision-making (SDM) similarly operationalizes a core principle of EBM by integrating information about available options, benefits, harms, and uncertainty with patients' values, preferences, and individual circumstances [15].

Together, these developments have strengthened both the trustworthiness and usability of evidence and the role of patient-centered decision-making within EBM.

Limitations of Contemporary EBM

Despite these advances, contemporary EBM faces several limitations. The rapidly expanding volume of research exceeds clinicians' capacity for continuous appraisal, while delays persist between evidence generation, synthesis, and implementation [6,16]. RCTs remain essential for estimating causal effects but may not adequately represent older adults, patients with multimorbidity or polypharmacy, and other populations encountered in routine practice. Observational studies, pragmatic trials, and routinely collected health data may therefore provide important complementary evidence, particularly when RCT evidence is limited or has restricted applicability [17,18].

Patient-centeredness also remains incompletely realized. Contemporary EBM requires greater emphasis on evidence that is usable in clinical practice and can be integrated with clinical context, professional expertise, shared decision-making, and individual patient preferences [16]. Moreover, conventional EBM has largely followed a linear pathway from research to evidence synthesis, recommendations, and practice, with less systematic use of data generated during routine care to produce new knowledge. Addressing this gap requires a more continuous connection between evidence generation and clinical practice, providing an important rationale for the development of next-generation EBM.

Concepts Shaping the Next Generation of EBM

The four approaches shaping next-generation EBM are complementary rather than competing and are summarized in Table 1.

Real-World Evidence

Real-world data (RWD) derived from electronic health records, administrative claims, registries, and digital health technologies can be used to generate evidence on the use, benefits, and harms of health interventions [18]. RWE can complement evidence from RCTs by providing information on broader and more heterogeneous populations, long-term outcomes, rare adverse events, and treatment patterns in routine clinical practice [17,18]. Advances in causal inference methods, including target trial emulation, may further improve the validity of causal estimates derived from observational data [19].

Living Evidence

Traditional systematic reviews and clinical practice guidelines may rapidly become outdated as new research accumulates. Living systematic reviews address this limitation by continually incorporating relevant new evidence as it becomes available [6]. Living guidelines extend this approach from evidence synthesis to recommendations, allowing guidance to be updated when new evidence is likely to change clinical recommendations [20]. Together, these approaches represent a shift from periodic publication toward continu-

Table 1. Complementary Approaches Shaping the Next Generation of Evidence-Based Medicine

Approach	Primary purpose	Main data sources	Temporal orientation	Contribution to EBM and key challenge
Real-world evidence	Evaluate effectiveness and safety in routine practice	Electronic health records, claims, registries, and patient-generated data	Historical or prospective	Improves generalizability beyond RCTs; vulnerable to confounding, bias, and variable data quality
Living evidence	Keep syntheses and guidance current	Continuous literature surveillance and newly available studies	Continuous or frequent updating	Reduces the delay between evidence generation and guidance; requires sustainable surveillance and updating workflows
Learning health systems	Convert practice data into knowledge and feed knowledge back into care	Clinical, operational, and patient-outcome data	Continuous feedback cycles	Connects evidence generation with implementation; depends on interoperability, governance, and organizational capacity
Precision medicine	Optimize decisions for individual patients	Genomic, biomarker, clinical, environmental, and life-style data	Individual-level, real-time, or repeated assessment	Extends decisions beyond average treatment effects; requires robust validation and attention to equity and access

EBM, evidence-based medicine; RCT, randomized controlled trial.

ous evidence surveillance, synthesis, and updating.

Learning Health Systems

A learning health system (LHS) integrates routinely generated clinical data, scientific evidence, and clinical experience through a continuous feedback process in which practice generates data, data generate knowledge, and knowledge informs subsequent practice [21]. In this model, clinical care becomes not only the endpoint of evidence implementation but also an important source of new evidence, creating a continuous cycle between evidence generation and practice [21].

Precision Medicine

Precision medicine seeks to account for individual variability in genes, environment, and lifestyle to improve disease prevention and treatment [22]. More broadly, genomic, biomarker, phenotypic, environmental, and behavioral information may be used to characterize patient heterogeneity and support more individualized clinical decisions [7,22]. When integrated with conventional research evidence and RWE, precision medicine may help move clinical decision-making beyond population-average treatment effects toward better estimation of outcomes for individual patients. Such personalization, however, requires adequate validation, predictive performance, and evidence of clinical utility before implementation in routine practice [7].

Evidence Ecosystems

Traditional EBM has largely followed a linear pathway from primary research through evidence synthesis and guideline development to clinical practice. A digital evidence ecosystem extends this model by continuously connecting evidence generation, synthesis, recommendations, implementation, outcomes, and new evidence [23].

Learning health systems emphasize the use of routinely generated clinical data to support continuous learning, whereas evidence ecosystems encompass a broader infrastructure for generating, synthesizing, updating, disseminating, and applying evidence [23]. Within such a system, living systematic reviews can inform living guidelines, recommendations can be incorporated into digital decision-support systems, and clinical outcomes can generate RWD that contribute to subsequent evidence generation and updating [20,21,23]. Achieving this continuous cycle requires trustworthy evidence synthesis, interoperable data and digital systems, and effective mechanisms for translating evidence into practice [23].

How Artificial Intelligence May Transform EBM

AI, including machine learning and large language models (LLMs), has the potential to support multiple stages of the EBM process. At Ask, AI may assist in structuring clinical questions; at Acquire, it can support literature searching, study identification, and screening; and at Appraise, AI tools are increasingly being evaluated for study classification, data extraction, and other components of evidence synthesis [24]. At Apply, AI may help integrate patient-level information with external evidence to support individualized risk and benefit estimation, while at Assess, it may facilitate analysis of clinical outcomes and identification of evidence gaps.

AI may be particularly valuable for living evidence systems, in which continuous surveillance and updating require substantial human resources. Automated or semi-automated tools can assist with identifying new studies, screening citations, extracting structured information, and determining when evidence syntheses may require updating [24].

However, AI-generated outputs cannot be assumed to be reliable. Potential problems include inaccurate or fabricated information, algorithmic bias, limited explainability, model instability, and challenges in reproducibility, transparency, and accountability. AI therefore reinforces rather than replaces core EBM principles: source verification, critical appraisal, assessment of uncertainty, transparency, and accountable human judgment. These potential applications, risks, and safeguards are summarized in Table 2.

Next-Generation EBM as a Digital Evidence Ecosystem

Next-generation EBM can be conceptualized as embedding the traditional principles and processes of EBM within a digitally connected and continuously learning evidence system [23]. Its evolution may be characterized by three major transitions: from static to living evidence, from population-average to more personalized evidence, and from a linear evidence pipeline to a learning evidence ecosystem [7,20,22,23].

In such a system, evidence syntheses and recommendations are continuously updated as new knowledge emerges; population-level evidence is integrated with individual risk, characteristics, and preferences; and data generated during routine clinical practice contribute to subsequent evidence generation. RWE, living evidence, LHSs, precision medicine, and AI can therefore be viewed not as alternatives to EBM but as complementary components of a continuous evidence cycle [18,20-24].

Important challenges remain, including data interoperability, privacy and governance, and the reliability, transparency, equity, and accountability of AI-supported systems. A trustworthy digital evidence ecosystem will therefore require

Table 2. Potential Applications and Safeguards of Artificial Intelligence Across the Evidence-Based Medicine Cycle

EBM stage	Potential applications of AI	Principal risks and safeguards
Ask	Structure clinical problems and generate focused PICO questions	Risk of poorly framed questions or loss of clinical context; require clinician review and explicit confirmation of assumptions
Acquire	Support semantic searching, citation screening, and study identification	Risk of missing relevant studies and limited search reproducibility; document search methods and validate recall against established approaches
Appraise	Assist with study-design identification, data extraction, and risk-of-bias assessment	Risk of hallucination and incorrect methodological judgments; use validated tools, human verification, and transparent audit trails
Apply	Support individualized risk prediction and clinical decision support	Risk of bias, inequity, poor calibration, and automation bias; require external validation, subgroup evaluation, explainability, and shared decision-making
Evaluate	Monitor outcomes, detect emerging signals, and identify evidence-practice gaps	Risk to privacy, accountability, and model stability; require governance, ongoing performance monitoring, and clear responsibility

AI, artificial intelligence; EBM, evidence-based medicine; PICO, population, intervention, comparison, and outcome.

integration across clinical epidemiology, evidence synthesis, health informatics, implementation science, patient participation, and data and AI governance. The goal is not to replace the foundational principles of EBM, but to enable those principles to operate within a more timely, adaptive, personalized, and continuously learning health care system [25].

Conclusion

EBM has evolved from individual appraisal of research into a broader framework encompassing evidence generation, synthesis, recommendations, implementation, and evaluation. Rapidly expanding evidence, implementation delays, limited generalizability, and the need for individualized decisions now require further evolution.

Next-generation EBM should integrate RWE, living evidence, LHSs, precision medicine, shared decision-making, and AI within a continuously learning evidence ecosystem. The AI era therefore does not signal the end of EBM; rather, it makes its central tasks—determining which evidence can be trusted, how uncertainty should be evaluated, and how evidence should be applied to individual patients—increasingly important.

Conflict of Interest

Soo Young Kim has been a member of the editorial board of the Journal of Evidence-Based Practice since 2025. However, he was not involved in the peer reviewer selection, evaluation, or decision process of this article. No other potential conflicts of interest relevant to this article were reported.

Funding

None.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Ethics Approval and Consent to Participate

Not applicable.

Authors' Contributions

Conceptualization: Kim SY. Methodology: Kim SY. Writing - original draft: Kim SY. Writing - review & editing: Kim SY.

Acknowledgments

None.

ORCID

Soo Young Kim, <https://orcid.org/0000-0002-3205-9408>

References

1. Evidence-Based Medicine Working Group. Evidence-based medicine: a new approach to teaching the practice of medicine. *JAMA* 1992; 268: 2420-5.
2. Djulbegovic B, Guyatt GH. Progress in evidence-based medicine: a quarter century on. *Lancet* 2017; 390: 415-23.
3. Sackett DL, Rosenberg WMC, Gray JAM, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. *BMJ* 1996; 312: 71-2.
4. Institute of Medicine. Finding what works in health care: standards for systematic reviews. Washington (DC), National Academies Press. 2011.
5. Institute of Medicine. Clinical practice guidelines we can trust.

- Washington (DC), National Academies Press. 2011.
6. Elliott JH, Turner T, Clavisi O, Thomas J, Higgins JPT, Mavergames C, et al. Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap. *PLoS Med* 2014; 11: e1001603.
 7. Beckmann JS, Lew D. Reconciling evidence-based medicine and precision medicine in the era of big data: challenges and opportunities. *Genome Med* 2016; 8: 134.
 8. Dawes M, Summerskill W, Glasziou P, Cartabellotta A, Martin J, Hopayian K, et al. Sicily statement on evidence-based practice. *BMC Med Educ* 2005; 5: 1.
 9. Albarqouni L, Hoffmann T, Straus S, Olsen NR, Young T, Ilic D, et al. Core competencies in evidence-based practice for health professionals: consensus statement based on a systematic review and Delphi survey. *JAMA Netw Open* 2018; 1: e180281.
 10. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008; 336: 924-6.
 11. Alonso-Coello P, Schunemann HJ, Moberg J, Brignardello-Pettersen R, Akl EA, Davoli M, et al. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: introduction. *BMJ* 2016; 353: i2016.
 12. Kim SY. Reporting guidelines. *Korean J Fam Med* 2009; 30: 62.
 13. De Angelis C, Drazen JM, Frizelle FA, Haug C, Hoey J, Horton R, et al. Clinical trial registration: a statement from the International Committee of Medical Journal Editors. *N Engl J Med* 2004; 351: 1250-1.
 14. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 2016; 3: 160018.
 15. Elwyn G, Frosch D, Thomson R, Joseph-Williams N, Lloyd A, Kinnersley P, et al. Shared decision making: a model for clinical practice. *J Gen Intern Med* 2012; 27: 1361-7.
 16. Greenhalgh T, Howick J, Maskrey N. Evidence based medicine: a movement in crisis? *BMJ* 2014; 348: g3725.
 17. Frieden TR. Evidence for health decision making-beyond randomized, controlled trials. *N Engl J Med* 2017; 377: 465-75.
 18. Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, et al. Real-world evidence-what is it and what can it tell us? *N Engl J Med* 2016; 375: 2293-7.
 19. Hernan MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *Am J Epidemiol* 2016; 183: 758-64.
 20. Akl EA, Meerpohl JJ, Elliott J, Kahale LA, Schunemann HJ; Living Systematic Review Network. Living systematic reviews: 4. Living guideline recommendations. *J Clin Epidemiol* 2017; 91: 47-53.
 21. Friedman CP, Rubin JC, Brown JS, Buntin MB, Corn M, Etheredge LM, et al. Toward a science of learning systems: a research agenda for the high-functioning Learning Health System. *J Am Med Inform Assoc* 2015; 22: 43-50.
 22. Collins FS, Varmus H. A new initiative on precision medicine. *N Engl J Med* 2015; 372: 793-5.
 23. Rajit D, Johnson A, Callander E, Teede H, Enticott J. Learning health systems and evidence ecosystems: a perspective on the future of evidence-based medicine and evidence-based guideline development. *Health Res Policy Syst* 2024; 22: 4.
 24. Harasgama S, Pearce H, Appel C, Loftus L, Painter H, Kuhn I, et al. Artificial intelligence tools for automating evidence synthesis: scoping review. *J Med Internet Res* 2026; 28: e81597.
 25. Institute of Medicine. Best care at lower cost: the path to continuously learning health care in America. Washington (DC), National Academies Press. 2013.

Sex- and age-specific associations between modifiable risk factors and incident Alzheimer's disease dementia: a nationwide cohort study in Korea

Hyun Jung Kim^{1,2}, Mookyoung Oh³

¹Cochrane Korea, Seoul, Korea

²Institute for Evidence-Based Medicine, College of Medicine, Korea University, Seoul, Korea

³Department of Internal Medicine, Soonchunhyang University Cheonan Hospital, Cheonan, Korea

Background: Modifiable risk factors account for a substantial proportion of dementia cases and represent important targets for prevention. However, previous studies have often focused on a limited range of risk factors and have rarely examined whether their associations with dementia differ according to sex and age at onset.

Objectives: To investigate sex- and age-specific associations between modifiable risk factors and incident Alzheimer's disease dementia using a nationwide population-based cohort in South Korea.

Methods: We conducted a retrospective cohort study using the National Health Information Database of the Korean National Health Insurance Service. Among individuals who participated in the National Cancer Screening Program in 2006 and underwent general health examinations in 2004, 2006, and 2008, 599,306 adults aged 40–79 years were included. Participants were followed from 2010 to 2019 after applying pre- and post-screening washout periods. Incident Alzheimer's disease dementia was defined using ICD-10 codes F00 or G30. Cox proportional hazards regression models were used to estimate hazard ratios (HRs) and 95% confidence intervals (CIs). Analyses were stratified by sex and age group to distinguish early-onset Alzheimer's disease dementia (EOAD) and late-onset Alzheimer's disease dementia (LOAD). During the 10-year follow-up period, dementia risk profiles differed substantially according to sex and age at onset. Lower educational attainment was consistently associated with increased dementia risk across all groups. Current smoking and physical inactivity were significant risk factors for dementia in both sexes, whereas underweight status was associated with an increased risk of LOAD. Among chronic conditions, diabetes, hypertension, and depression elevated dementia risk, with depression demonstrating the strongest association, particularly for EOAD (males: HR = 3.61, 95% CI: 2.63–4.96; females: HR = 1.83, 95% CI: 1.46–2.29). Sensory and functional impairments, including hearing loss, visual loss, physical disability, and traumatic brain injury, were also associated with increased dementia risk, although the magnitude and significance of these associations varied by sex and age group. The predictive performance of the models was acceptable, with C-statistics ranging from 0.6784 to 0.7803.

Conclusions: The associations between modifiable risk factors and incident Alzheimer's disease dementia exhibited distinct sex- and age-specific patterns. These findings highlight the importance of tailored dementia prevention strategies that account for differences in demographic characteristics and risk factor profiles. Targeted interventions addressing modifiable risk factors may contribute to reducing the future burden of dementia at both individual and population levels.

Keywords: Alzheimer's disease dementia; modifiable risk factors; Cox proportional hazards regression

Received: June 5, 2026; **Accepted:** August 25, 2026

Corresponding author: Mookyoung Oh

Email: somanghana@gmail.com

© 2026 Korean Society of Evidence-Based Medicine

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Introduction

Dementia is a syndrome encompassing common symptoms such as cognitive decline and behavioral changes caused by brain diseases [1,2]. In South Korea, the prevalence of dementia among adults aged 65 years and older reached 9.25% in 2023 [3]. The national disease burden of dementia, measured in disability-adjusted life years (DALYs), doubled between 2000 and 2019, representing the steepest increase among major chronic diseases [4]. With rapid population aging, dementia is expected to impose an even greater public health burden in the coming decades [5].

When classified according to reversibility, irreversible dementia accounts for approximately 60–95% of all cases. Importantly, neuropathological changes associated with dementia are known to begin 15–25 years before clinical diagnosis [6]. Moreover, clinical diagnosis is typically made only after cognitive impairment interferes with daily functioning, at which point substantial neuropathological damage has already occurred. From an epidemiological perspective, this long preclinical phase provides a critical window for primary prevention. Therefore, identifying individuals at high risk before clinical onset and implementing early preventive interventions represent key public health priorities.

Dementia risk factors are heterogeneous and multifactorial, encompassing health behaviors, cardiometabolic conditions, psychosocial factors, and functional impairments [7]. Given the irreversible nature of dementia and its increasing societal burden, growing attention has been directed toward modifiable risk factors from both prevention and management perspectives. Although earlier epidemiological evidence was limited, advances in etiologic classification and pathophysiological understanding have led to a growing body of observational research on dementia risk factors. Accordingly, numerous cohort studies have investigated the association between modifiable risk factors—such as smoking, alcohol consumption, obesity, and hypertension—and cognitive decline, typically adjusting for sociodemographic covariates [8–15]. Notably, recent studies using population attributable fraction (PAF) analyses have estimated that approximately 45% of dementia cases are potentially preventable by addressing these modifiable risk factors, providing a powerful rationale for population-level intervention strategies [16].

However, existing literature has several limitations. Previous studies have predominantly evaluated dementia risk by focusing on a narrow range of conventional risk factors, such as blood pressure and total cholesterol [17,18]. Crucial comorbidities and sensory or physical impairments, including depression, visual loss, and hearing loss, have often been

excluded from comprehensive analyses. Consequently, the distinct impacts of various risk factor profiles on dementia onset remain poorly understood. This gap in the literature primarily stems from the lack of data sources that comprehensively capture socioeconomic factors, health behaviors, chronic diseases, and a wide range of dementia-related risk factors, as well as study design constraints, including the scarcity of comprehensive longitudinal data. Furthermore, while PAF metrics are invaluable for guiding public health policies to improve population-level health, they fundamentally quantify aggregate population risk and fail to evaluate individual-level dementia risk.

To address these limitations, this study aimed to examine sex- and age-specific associations between modifiable risk factors and incident Alzheimer's disease dementia using the National Health Information Database (NHID) managed by the National Health Insurance Service (NHIS) in South Korea. This nationwide database, initiated in 2000, accumulates extensive health screening data for Korean adults aged 40 years and older, capturing diverse health metrics including self-reported lifestyle questionnaires, physical examinations, and laboratory findings. To ensure robust and stable quantification of health behaviors, chronic comorbidities, and functional impairments, we selected a well-defined cohort of individuals who underwent three consecutive biennial health screenings over a 6-year baseline period. Utilizing a subsequent 10-year longitudinal follow-up, we estimated the long-term risk of Alzheimer's disease dementia attributable to specific modifiable risk factors, with detailed stratification by sex and age groups. Notably, by linkable integration with the National Cancer Screening Program database, we successfully ascertained individual educational attainment—a well-established proxy for cognitive reserve and a key confounding factor [19,20]—thereby allowing a comprehensive array of dementia risk factors to be included in our analysis.

Methods

Data source and study population

This nationwide retrospective cohort study was conducted using customized data provided by the National Health Insurance Service (NHIS) of Korea, which covers the entire Korean population. The NHIS database includes eligibility records, health screening results, healthcare utilization claims, and mortality information.

Among 3,529,471 individuals who participated in the 2006 National Cancer Screening Program, we identified 1,170,432 individuals who additionally underwent general health examinations in 2004, 2006, and 2008. To ensure temporal

ordering between exposure changes and dementia onset, we applied both pre-screening and post-screening wash-out periods. To exclude prevalent dementia cases, individuals who diagnosed with dementia before January 1, 2009 were excluded ($n = 6,004$). Second, those aged <40 or ≥ 80 years at baseline in 2009 ($n = 18,876$) were excluded. Finally, participants with missing values in any screening variables were excluded ($n = 546,246$), resulting in a final analytic cohort of 599,306 individuals.

Follow-up and outcome ascertainment

The period from 2002 to the first health examination in 2004 was defined as the pre-screening wash-out period to exclude individuals with prior dementia diagnosis. To account for the latency between risk factor modification and dementia onset, a one-year post-screening washout period (2009), extending from the date of the participant's last health screening exam-

ination, was applied. Dementia cases diagnosed during this period were excluded to minimize reverse causation (Fig. 1). Participants were followed from January 1, 2009, until the first diagnosis of Alzheimer's disease dementia, death, or December 31, 2019, whichever occurred first. Alzheimer's disease dementia was defined using ICD-10 codes F00 or G30 recorded as primary or secondary diagnoses; additional operational criteria should be specified if prescription or repeated-claim information was used (Fig. 2).

Definition of risk factors and outcome ascertainment

Health behavior factors previously identified as risk factors for dementia—including smoking, high-risk alcohol consumption, physical inactivity, and obesity—were included in the analysis. In addition, chronic conditions such as diabetes mellitus, hypertension, and depression, as well as disability including hearing loss, visual loss, and physical disability.

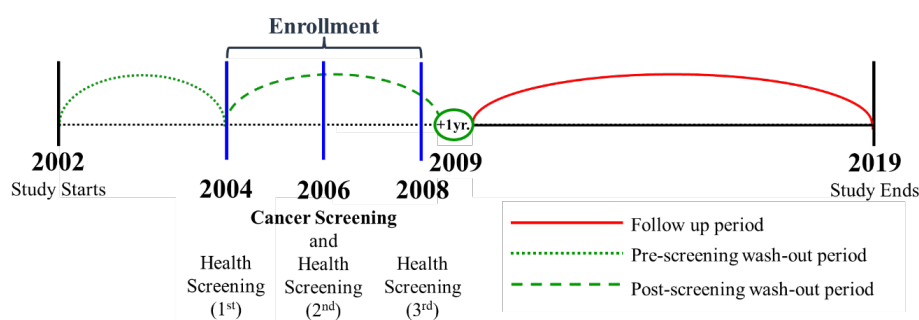


Fig. 1. Framework of the follow up period.

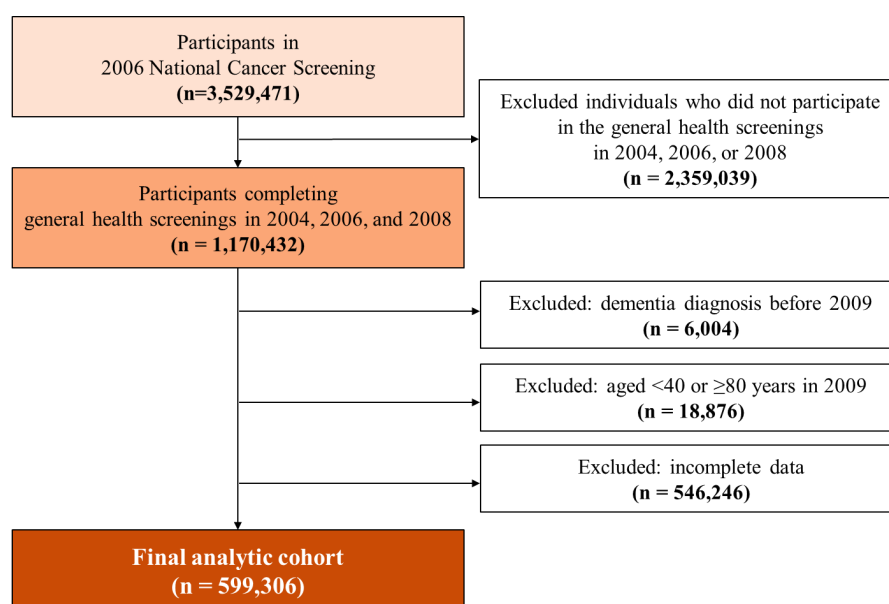


Fig. 2. Process of study sample selection.

And traumatic brain injury was incorporated. Furthermore, sociodemographic variables, including sex, age, educational attainment and health insurance premium quantile as income, were added to the model. In total, 16 risk factors were included in the final analysis.

Incident Alzheimer's disease dementia was ascertained between 2010 and 2019 using claims-based ICD-10 diagnosis codes F00 or G30 listed as a primary or secondary diagnosis.

Some dementia cases cannot be unequivocally attributed to a single etiologic subtype. In particular, in patients with Alzheimer's disease who exhibit concomitant cerebrovascular pathology, the National Institute of Neurological Disorders and Stroke–Association Internationale pour la Recherche et l'Enseignement en Neurosciences (NINDS-AIREN) criteria recommend classifying such cases as Alzheimer's disease with cerebrovascular disease (AD with CVD), rather than applying the nonspecific term 'mixed dementia,' which has been widely used but lacks standardized diagnostic definition [21]. In accordance with this recommendation, when individuals met diagnostic criteria for both Alzheimer's disease dementia and vascular dementia, priority was assigned to Alzheimer's disease dementia in the present study to ensure mutually exclusive subtype classification.

Statistical analysis and development of the risk prediction model

Cox proportional hazards regression models were used to estimate hazard ratios (HRs) and 95% confidence intervals (CIs) for incident dementia. All 16 risk factors were simultaneously included in multivariable models to estimate independent associations.

The study population was stratified by sex (males and females) and subsequently by age at baseline in 2009 (<55 years and ≥ 55 years). The baseline year of 2009 was specifically designated so that participants would reach the ages of either <65 or ≥ 65 years by the end of the follow-up period in 2019. This stratification strategy was implemented to adjust for the confounding effects of age and to effectively distinguish between developing before age 65 (Early Onset Alzheimer's Dementia; EOAD) and developing at or after age 65 (Late-Onset Alzheimer's Dementia; LOAD) during the follow-up period [22].

Results

General characteristics

The final analytic cohort included 599,306 participants. Baseline characteristics differed by sex in sociodemographic factors and health behaviors (Table 1).

Table 1. General Characteristics of Study Population

Variables	Male (N=263,273)		Female (N=336,033)		SMD*
	N	%	N	%	
Mean Age('09)	56.27±9.2		55.94±8.6		0.03804
Age group(years old)					
40~44	18,658	7.1	16,976	5.1	0.15055
45~49	56,331	21.4	79,263	23.6	
50~54	52,888	20.1	63,441	18.9	
55~59	50,250	19.1	76,221	22.7	
60~64	30,577	11.6	39,967	11.9	
65~69	24,320	9.2	30,081	9.0	
70~74	18,589	7.1	19,115	5.7	
≥ 75	11,660	4.4	10,969	3.3	
Income					
Medical Aid & 1Q (low)	27,972	10.6	67,708	20.1	0.28105
2Q	38,391	14.6	53,447	15.9	
3Q	66,134	25.1	74,613	22.2	
4Q (high)	130,776	49.7	140,265	41.7	
Education					
Unschool	7,172	2.7	24,143	7.2	0.45373
Elementary School	29,903	11.4	70,748	21.1	
Middle School	36,487	13.9	61,860	18.4	
High School	88,901	33.8	103,719	30.9	
College student or higher	100,810	38.3	75,563	22.5	
Spouse					
Yes	249,352	94.7	290,990	86.6	0.28116
No	13,921	5.3	45,043	13.4	
Smoking					
None smoker	121,966	46.3	327,778	97.5	1.38728
Former smoker	51,117	19.4	3,287	1.0	
Current smoker	90,190	34.3	4,968	1.5	
Alcohol					
Non-drinker	90,448	34.4	275,451	82.0	1.10817
Moderate drinker	131,841	50.1	51,283	15.3	
Heavy drinker	40,984	15.6	9299	2.8	
Physical Activity					
< 1 time/week	112,091	42.6	183,762	54.7	0.24408
≥ 1 times/week	151,182	57.4	152,271	45.3	
BMI					
Normal weight & Obesity	258,231	98.1	327,666	97.5	0.03917
Underweight	5,042	1.9	8,367	2.5	
Diabetes					
Normal	229,624	87.2	305,593	90.9	0.11956
Diabetes	33,649	12.8	30,440	9.1	
Hypertension					
Normal	156,536	59.5	228,751	68.1	0.17998
Hypertension	106,737	40.5	107,282	31.9	

(Continued to the next column)

Table 1. Continued

Variables	Male (N=263,273)		Female (N=336,033)		SMD*
	N	%	N	%	
Dyslipidemia					
Normal	150,321	57.1	184,259	54.8	0.04561
Dyslipidemia	112,952	42.9	151,774	45.2	
Depression					
Normal	255,466	97.0	316,809	94.3	0.13550
Depression	7,807	3.0	19,224	5.7	
TBI					
None	261,480	99.3	334,820	99.6	0.04447
TBI	1,793	0.7	1213	0.4	
Visual loss					
None	254,115	96.5	331,387	98.6	0.13642
Visual loss	9,158	3.5	4,646	1.4	
Hearing loss					
None	261,793	99.4	335,232	99.8	0.05130
Hearing loss	1,480	0.6	801	0.2	
Physical Disability					
None	262,174	99.6	335,467	99.8	0.04609
Physical Disability	1,099	0.4	566	0.2	

SMD, standard mean difference; BMI, body mass index; DM, diabetes mellitus; TBI, traumatic brain injury.

Of the 599,306 total participants, 263,273 were male and 336,033 were female. At baseline in 2009, the mean age was 56.27 ± 9.2 years for males and 55.94 ± 8.6 years for females. In the 5-year age group analysis, the 45~49 cohort was the most prevalent in both sexes (males: 21.4%, females: 23.6%), whereas the ≥ 75 cohort was the least prevalent (males: 4.4%, females: 3.3%).

Regarding socioeconomic status, females were characterized by lower income and lower educational attainment than males. The proportion in the lowest income quantile (Medical Aid & 1Q) was 10.6% for males versus 20.1% for females, while the highest quantile (4Q) accounted for 49.7% of males and 41.7% of females. Similarly, the prevalence of being unschooled was lower in males (2.7% vs. 7.2%), whereas high school graduation was higher (33.8% vs. 30.9%). For marital status, the proportion of those who were divorced, widowed, or separated was substantially lower in males than in females (3.2% vs. 11.1%).

Health behavior profiles exhibited prominent sex-specific disparities. The current smoking rate was 34.3% among males compared with only 1.5% among females, while 97.5% of females were non-smokers. Heavy drinking was observed in 15.6% of males and 2.8% of females. In addition, 57.4% of males engaged in physical activity ≥ 1 times/week, compared with 45.3% of females. The prevalence of underweight was

1.9% in males and 2.5% in females.

Regarding chronic comorbidities, diabetes (12.8% vs. 9.1%) and hypertension (40.5% vs. 31.9%) were more prevalent in males, whereas dyslipidemia (42.9% vs. 45.2%) and depression (3.0% vs. 5.7%) were more frequent in females. For injuries and sensory/physical impairments, males showed a higher prevalence across all categories: traumatic brain injury (0.7% vs. 0.4%), visual loss (3.5% vs. 1.4%), hearing loss (0.6% vs. 0.2%), and physical disability (0.4% vs. 0.2%).

Overall, substantial differences between sexes were observed primarily in health behaviors, with standardized mean differences (SMDs) exceeding 0.8 (smoking: SMD = 1.39; alcohol: SMD = 1.11). In contrast, differences in age, income, education, marital status, physical activity, chronic diseases, and impairments were minimal, with all SMDs remaining below 0.5.

Risk of alzheimer's dementia

Males

Dementia risk factor profiles in males showed distinct variations between early-onset dementia (EOAD) and late-onset dementia (LOAD) (Table 2). Sociodemographically, the risk of dementia increased sequentially for every 5-year age increment in both subtypes. Lower income (Medical Aid & 1Q) significantly elevated the risk of EOAD (hazard ratio [HR] = 1.54, 95% confidence interval [CI]: 1.13–2.10), whereas lower educational attainment was consistently and significantly associated with an increased risk of LOAD across all education levels (e.g., Unschooled: HR = 1.52, 95% CI: 1.40–1.66). Although individuals without a spouse (divorced/widowed/separated) tended to increase EOAD risk, it was not statistically significant (HR = 1.11, $P = 0.56$).

Among health behaviors, current smoking (HR = 1.19, 95% CI: 1.13–1.24) and physical inactivity (HR = 1.12, 95% CI: 1.08–1.17) significantly increased LOAD risk, but not EOAD risk. Being underweight did not reach statistical significance for either EOAD (HR = 1.77, $P = 0.06$) or LOAD (HR = 1.14, $P = 0.02$).

For chronic comorbidities and impairments, both diabetes and depression significantly increased dementia risk across all age groups. Notably, depression was a remarkably potent risk factor for EOAD (HR = 3.61, 95% CI: 2.63–4.96) and remained significant for LOAD (HR = 1.52, 95% CI: 1.41–1.64). Traumatic brain injury (TBI) significantly elevated the risk of EOAD (HR = 2.52, 95% CI: 1.30–4.90) and LOAD (HR = 1.54, 95% CI 1.33–1.79). Conversely, hearing loss and physical disability were significant risk factors for LOAD. In contrast, visual loss was not significantly associated with dementia risk

in males. The C-statistics for the male models were 0.6784 for EOAD and 0.7803 for LOAD, respectively.

Females

In females, risk profiles also demonstrated substantial differences between EOAD and LOAD (Table 2). For sociodemographic factors, a 5-year age increment consistently doubled the risk of both subtypes. Lower education levels significantly increased the risk of both conditions, with a more pronounced, graded relationship in LOAD (Unschooling: HR = 1.84, 95% CI: 1.69–2.00). Notably, the absence of a spouse was significantly associated with a decreased risk of LOAD (HR = 0.95, 95% CI: 0.92–0.98).

Regarding health behaviors, current smoking increased the risk of both EOAD (HR = 1.46, 95% CI: 0.99–2.14) and LOAD (HR = 1.26, 95% CI: 1.14–1.38). Physical inactivity significantly elevated the risk for both EOAD (HR = 1.22, 95% CI: 1.08–1.38) and LOAD (HR = 1.12, 95% CI: 1.08–1.15). Being underweight significantly increased LOAD risk (HR = 1.25, 95% CI: 1.14–1.38), whereas obesity was protective against LOAD (HR = 0.91, 95% CI: 0.84–0.99).

Among comorbidities and impairments, diabetes and depression significantly elevated the risk for both EOAD and LOAD, with depression showing a significant risk in EOAD (HR = 1.83, 95% CI: 1.46–2.29). For disabilities, physical disability (HR = 2.34, 95% CI: 1.56–3.51) and hearing loss (HR = 2.93, 95% CI: 1.10–7.84) were significant risk factors uniquely for EOAD. Conversely, TBI (HR = 1.25, 95% CI: 1.05–1.50), physical disability (HR = 1.19, 95% CI: 1.09–1.30), and visual loss (HR = 1.27, 95% CI: 1.05–1.54) were associated with an increased risk for LOAD. The C-statistics for the female models were 0.6857 for EOAD and 0.7772 for LOAD, respectively.

Discussion

The risk of dementia associated with modifiable risk factors exhibited distinct sex- and age-specific characteristics. Regarding socioeconomic status, lower educational attainment and lower income levels were robustly associated with an increased risk of dementia across all sexes and age groups. Furthermore, while cohabitation with a spouse is conventionally reported as a protective factor against dementia across diverse cultural backgrounds [23], our study revealed a notable sex-based divergence: the absence of a spouse elevated dementia risk in males but lowered it in females. This discrepancy suggests that a spouse functions differently between sexes as a mediator of social contact—such as connections with family, friends, and neighbors. It may also reflect sex differences in marital satisfaction, spousal caregiving burdens,

and marital conflict.

In terms of health behaviors, current smoking and physical inactivity were associated with a significantly increased risk of dementia in both sexes. With respect to body mass index (BMI), numerous studies have reported that obesity ($\text{BMI} \geq 30 \text{ kg/m}^2$) is associated with an elevated risk of dementia [24]. In contrast, our findings identified underweight status ($\text{BMI} < 18.5 \text{ kg/m}^2$) as a significant risk factor for dementia [25,26]. This finding is consistent with recent domestic and international cohort studies demonstrating the adverse health consequences and increased dementia risk associated with being underweight. Notably, these studies have suggested that the association becomes more pronounced with advancing age.

Among chronic comorbidities, diabetes, hypertension, and depression consistently increased dementia risk in both sexes. Notably, depression emerged as the most potent risk factor, exhibiting the highest hazard ratio among all examined variables. Furthermore, the presence of traumatic brain injury (TBI), visual loss, hearing loss, or physical disability consistently elevated dementia risk regardless of sex. Rather than the disability itself directly driving the pathogenesis, this association is likely mediated by disability-induced reductions in physical activity and increased social isolation. For hearing loss, psychosocial factors such as loneliness, depression, and social isolation caused by hearing impairment have been continuously implicated [27]. Mechanistically, this is primarily explained by a neuropathological interplay: decreased external sensory stimulation leads to a reduction in cognitive reserve, while the compensatory need for increased cognitive resources during listening tasks further taxes the brain.

This study has several limitations. First, the broad definition of dementia used in this study may have resulted in the inclusion of individuals who were overdiagnosed or misclassified as having dementia. Differences in the definition of dementia cases across studies have led to substantial variation in estimates of dementia prevalence and incidence [28,29]. Even when examining all-cause dementia, researchers have applied different diagnostic codes, varied the inclusion of primary and secondary diagnoses, and adopted different healthcare utilization criteria, such as requiring only outpatient visits or additionally considering dementia-related medication use. In this study, individuals were classified as having Alzheimer's disease dementia if they had at least one outpatient visit with a diagnosis of Alzheimer's disease dementia (ICD-10 codes F00 or G30) during the follow-up period. Both primary and secondary diagnoses were considered without distinction. This broad case definition was adopted

Table 2. Age-Specific Risk of Alzheimer's Dementia by Risk Factors

Variables	Male					Female				
	HR	P	95% CI	HR	LOAD	HR	P	95% CI	HR	LOAD
Age group(years old)										
40~44	1.00	-	ref.			1.00	-	ref.		
45~49	2.01	0.00	(1.33~3.04)			2.42	0.00	(1.63~3.58)		
50~54	3.84	0.00	(2.56~5.76)			5.06	0.00	(3.42~7.49)		
55~59				1.00	-			ref.	1.00	-
60~64				2.53	0.00			(2.33~2.75)	2.24	0.00
65~69				5.39	0.00			(4.99~5.82)	4.89	0.00
70~74				10.3	0.00			(9.56~11.10)	9.14	0.00
≥ 75				17.26	0.00			(15.99~18.63)	14.3	0.00
Income										
Medical Aid & 1Q	1.54	0.01	(1.13~2.10)	1.07	0.02	0.77	0.00	(0.64~0.92)	1.1	0.00
2Q	1.13	0.39	(0.85~1.50)	1.12	0.00	1.08	0.39	(0.90~1.30)	1.13	0.00
3Q	1.12	0.31	(0.90~1.39)	1.09	0.00	1.18	0.04	(1.01~1.39)	1.05	0.02
4Q	1.00	-	ref.	1.00	-	1.00	-	ref.	1.00	-
Education										
Unschooling	1.49	0.23	(0.77~2.88)	1.52	0.00	1.43	0.04	(1.01~2.03)	1.84	0.00
Elementary School	2.06	0.00	(1.48~2.87)	1.48	0.00	2.01	0.00	(1.62~2.50)	1.64	0.00
Middle School	1.47	0.01	(1.10~1.96)	1.26	0.00	1.88	0.00	(1.55~2.27)	1.41	0.00
High School	1.02	0.84	(0.83~1.26)	1.19	0.00	1.31	0.00	(1.10~1.55)	1.29	0.00
College student or higher	1.00	-	ref.	1.00	-	1.00	-	ref.	1.00	-
Spouse										
Yes	1.00	-	ref.	1.00	-	1.00	-	ref.	1.00	-
No	1.11	0.56	(0.78~1.58)	1.05	0.23	1.12	0.24	(0.92~1.37)	0.95	0.00
Smoking										
None smoker	1.00	-	ref.	1.00	-	1.00	-	ref.	1.00	-
Former smoker	0.91	0.47	(0.70~1.18)	1.05	0.06	0.96	0.90	(0.53~1.75)	1.26	0.00
Current smoker	1.19	0.09	(0.97~1.45)	1.19	0.00	1.46	0.06	(0.99~2.14)	1.26	0.00
Alcohol										
Non-drinker	1.00	-	ref.	1.00	-	1.00	-	ref.	1.00	-
Moderate drinker	0.81	0.04	(0.66~0.99)	0.89	0.00	0.99	0.87	(0.84~1.15)	0.99	0.64
Heavy drinker	0.89	0.41	(0.67~1.18)	1.01	0.84	1.31	0.06	(0.99~1.73)	1.08	0.17
Physical Activity										
< 1 times/week	0.97	0.74	(0.81~1.16)	1.12	0.00	1.22	0.00	(1.08~1.38)	1.12	0.00
≥ 1 times/week	1.00	-	ref.	1.00	-	1.00	-	ref.	1.00	-

(Continued to the next page)

Table 2. Continued

Variables	Male						Female					
	EOAD			LOAD			EOAD			LOAD		
	HR	P	95% CI	HR	P	95% CI	HR	P	95% CI	HR	P	95% CI
BMI†												
Underweight	1.77	0.06	(0.97-3.23)	1.14	0.02	(1.02-1.27)	0.92	0.70	(0.61-1.39)	1.25	0.00	(1.14-1.38)
Normal weight	1.00	-	ref.	1.00	-	ref.	1.00	-	ref.	1.00	-	ref.
Obesity	0.73	0.36	(0.38-1.42)	1.03	0.70	(0.89-1.20)	0.86	0.49	(0.55-1.33)	0.91	0.02	(0.84-0.99)
Diabetes												
Normal	1.00	-	ref.	1.00	-	ref.	1.00	-	ref.	1.00	-	ref.
Diabetes	1.79	0.00	(1.40-2.29)	1.38	0.00	(1.33-1.45)	1.51	0.00	(1.19-1.93)	1.35	0.00	(1.31-1.40)
Hypertension												
Normal	1.00	-	ref.	1.00	-	ref.	1.00	-	ref.	1.00	-	ref.
Hypertension	1.08	0.41	(0.90-1.30)	1.07	0.00	(1.03-1.12)	1.02	0.75	(0.88-1.19)	1.06	0.00	(1.03-1.09)
Dyslipidemia												
Normal	1.00	-	ref.	1.00	-	ref.	1.00	-	ref.	1.00	-	ref.
Dyslipidemia	1.10	0.31	(0.92-1.31)	1.02	0.32	(0.98-1.06)	1.07	0.26	(0.95-1.21)	1.01	0.41	(0.98-1.05)
Depression												
Normal	1.00	-	ref.	1.00	-	ref.	1.00	-	ref.	1.00	-	ref.
Depression	3.61	0.00	(2.63-4.96)	1.52	0.00	(1.41-1.64)	1.83	0.00	(1.46-2.29)	1.56	0.00	(1.49-1.63)
TBI												
None	1.00	-	ref.	1.00	-	ref.	1.00	-	ref.	1.00	-	ref.
TBI	2.52	0.01	(1.30-4.90)	1.54	0.00	(1.33-1.79)	0.84	0.76	(0.27-2.61)	1.25	0.01	(1.05-1.50)
Physical Disability												
None	1.00	-	ref.	1.00	-	ref.	1.00	-	ref.	1.00	-	ref.
Physical Disability	1.44	0.09	(0.95-2.18)	1.2	0.00	(1.11-1.30)	2.34	0.00	(1.56-3.51)	1.19	0.00	(1.09-1.30)
Visual loss												
None	1.00	-	ref.	1.00	-	ref.	1.00	-	ref.	1.00	-	ref.
Visual loss	0.99	0.99	(0.25-3.99)	1.14	0.14	(0.96-1.36)	1.48	0.58	(0.37-5.92)	1.27	0.01	(1.05-1.54)
Hearing loss												
None	1.00	-	ref.	1.00	-	ref.	1.00	-	ref.	1.00	-	ref.
Hearing loss	1.57	0.53	(0.39-6.31)	1.34	0.00	(1.13-1.59)	2.93	0.03	(1.10-7.84)	1.21	0.11	(0.96-1.54)
C-statistics		0.6784			0.7803			0.6857			0.7772	

for several reasons. First, the primary objective of this study was to investigate the association between dementia risk factors and subsequent dementia incidence. Therefore, it was considered appropriate to include not only patients with Alzheimer's disease dementia but also individuals at high risk of dementia, such as those with mild cognitive impairment. Second, continued healthcare utilization may not occur even after a dementia diagnosis because of factors such as therapeutic nihilism regarding the perceived untreatability of dementia and the social stigma associated with a dementia diagnosis, both of which may discourage ongoing medical care [30,31]. Third, given the irreversible neuropathological changes characteristic of Alzheimer's disease dementia, a single healthcare encounter with an Alzheimer's disease dementia diagnosis is likely to represent a true dementia case. Nevertheless, the use of a broad diagnostic definition may have increased the possibility of outcome misclassification, and the findings should therefore be interpreted with caution. Future studies employing more stringent diagnostic criteria, including repeated healthcare encounters, medication records, or clinically validated diagnoses, are warranted to confirm the robustness of the observed associations.

A second limitation of this study relates to the age classification strategy. Participants were categorized as either younger than 55 years or 55 years and older at baseline (2009), the start of follow-up. This classification was intended to distinguish dementia cases that occurred during the follow-up period into early-onset Alzheimer's disease (EOAD) and late-onset Alzheimer's disease (LOAD), based on whether participants would be younger than 65 years or 65 years and older, respectively, by the end of follow-up in 2019.

This dichotomization at 55 years may have introduced some degree of misclassification, particularly for individuals near the cutoff point, who could have been assigned to different age groups despite having similar underlying risk profiles. Consequently, the effect of age on dementia risk may have been amplified, whereas the relative contributions of other risk factors may have been attenuated. Nevertheless, stratifying participants into EOAD and LOAD groups allowed us to better capture the potentially nonlinear relationship between age and dementia risk, improved adherence to the proportional hazards assumption, and enhanced the clinical interpretability and applicability of the findings. Therefore, we considered the distinction between EOAD and LOAD to be appropriate for addressing the objectives of this study.

Conclusions

The risk of dementia associated with specific risk factors ex-

hibited distinct sex- and age-specific characteristics. Consequently, tailored approaches encompassing both individualized prevention planning and population-level risk reduction strategies are warranted to effectively mitigate the growing burden of dementia.

Conflict of Interest

Hyun Jung Kim has been a editor of Journal of Evidence-Based Practice since 2025. However, she was not involved in the peer reviewer selection, evaluation, or decision process of this article. No other potential conflicts of interest relevant to this article were reported.

Funding

This work was not directly funded.

Data Availability Statement

The data that support the findings of this study are available from the Korean National Health Insurance Service (NHIS) but restrictions apply to the availability of these data, which were used under license for the current study. Data are available from the authors upon reasonable request and with permission of NHIS.

Ethics Approval and Consent to Participate

The study protocol was approved by the Institutional Review Board (IRB No. E-2411-131-1590) and the National Health Information Data Review Committee of the National Health Insurance Service (Research Management No. REQ2025010906-006). The requirement for informed consent was waived because all data were fully anonymized prior to analysis.

Authors' Contributions

Conceptualization: Mookyoung Oh, Hyun Jung Kim. Methodology: Mookyoung Oh, Hyun Jung Kim. Software: Mookyoung Oh. Investigation: Hyun Jung Kim. Original draft: Mookyoung Oh. Writing—review and editing: Mookyoung Oh, Hyun Jung Kim. Supervision: Hyun Jung Kim.

Acknowledgments

We thank the Korea NHIS for providing the database used in this study.

ORCID

Hyun Jung Kim, <https://orcid.org/0000-0003-2018-2385>

Mookyung Oh, <https://orcid.org/0000-0002-2011-5708>

References

- World Health Organization. The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines. Geneva, WHO. 1992.
- Morris JC. The Clinical Dementia Rating (CDR): current version and scoring rules. *Neurology* 1993; 43: 2412-4.
- Ministry of Health and Welfare, National Institute of Dementia. Korean nationwide dementia epidemiology and status survey 2023. Sejong, Ministry of Health and Welfare. 2025.
- Avan A, Hachinski V. Global, regional, and national trends of dementia incidence and risk factors, 1990-2019: A Global Burden of Disease study. *Alzheimers Dement* 2023; 19: 1281-91.
- Stephan BCM, Birdi R, Tang EYH, Cosco TD, Donini LM, Licher S, et al. Secular Trends in Dementia Prevalence and Incidence Worldwide: A Systematic Review. *J Alzheimers Dis* 2018; 66: 653-80.
- Gale SA, Acar D, Daffner KR. Dementia. *Am J Med* 2018; 131: 1161-9.
- Dening T, Sandilyan MB. Dementia: definitions and types. *Nurs Stand* 2015; 29: 37-42.
- Hwang PH, Ang TFA, De Anda-Duran I, Liu X, Liu Y, Gurnani A, et al. Examination of potentially modifiable dementia risk factors across the adult life course: The Framingham Heart Study. *Alzheimers Dement* 2023; 19: 2975-83.
- Gottesman RE, Albert MS, Alonso A, Coker LH, Coresh J, Davis SM, et al. Associations Between Midlife Vascular Risk Factors and 25-Year Incident Dementia in the Atherosclerosis Risk in Communities (ARIC) Cohort. *JAMA Neurol* 2017; 74: 1246-54.
- Raggi M, Dugravot A, Valeri L, Machado-Fragua MD, Dumurgier J, Kivimaki M, et al. Contribution of smoking towards the association between socioeconomic position and dementia: 32-year follow-up of the Whitehall II prospective cohort study. *Lancet Reg Health Eur* 2022; 23: 100516.
- Wu S, Xu Y, Zheng R, Lu J, Li M, Chen L, et al. Hypertension Defined by 2017 ACC/AHA Guideline, Ideal Cardiovascular Health Metrics, and Risk of Cardiovascular Disease: A Nationwide Prospective Cohort Study. *Lancet Reg Health West Pac* 2022; 20: 100350.
- Peters R, Booth A, Rockwood K, Peters J, D'Este C, Anstey KJ. Combining modifiable risk factors and risk of dementia: a systematic review and meta-analysis. *BMJ Open* 2019; 9: e022846.
- Elias MF, Elias PK, Sullivan LM, Wolf PA, D'Agostino RB. Lower cognitive function in the presence of obesity and hypertension: the Framingham heart study. *Int J Obes Relat Metab Disord* 2003; 27: 260-8.
- Hagger-Johnson G, Sabia S, Brunner EJ, Shipley M, Bobak M, Marmot M, et al. Combined impact of smoking and heavy alcohol use on cognitive decline in early old age: Whitehall II prospective cohort study. *Br J Psychiatry* 2013; 203: 120-5.
- Whitmer RA, Sidney S, Selby J, Johnston SC, Yaffe K. Midlife cardiovascular risk factors and risk of dementia in late life. *Neurology* 2005; 64: 277-81.
- Livingston G, Huntley J, Liu KY, Costafreda SG, Selbæk G, Aladi S, et al. Dementia prevention, intervention, and care: 2024 report of the Lancet standing Commission. *Lancet* 2024; 404: 572-628.
- Kimm H, Lee PH, Shin YJ, Park KS, Jo J, Lee Y, et al. Mid-life and late-life vascular risk factors and dementia in Korean men and women. *Arch Gerontol Geriatr* 2011; 52: e117-22.
- Son SK. Analysis of the risk of dementia associated with diabetes mellitus in the Korean elderly population [master's thesis]. Seoul, Korea University; 2020.
- Sharp ES, Gatz M. Relationship between education and dementia: an updated systematic review. *Alzheimer Dis Assoc Disord* 2011; 25: 289-304.
- Norton S, Matthews FE, Barnes DE, Yaffe K, Brayne C. Potential for primary prevention of Alzheimer's disease: an analysis of population-based data. *Lancet Neurol* 2014; 13: 788-94.
- Román GC, Tatemichi TK, Erkinjuntti T, Cummings JL, Masdeu JC, Garcia JH, et al. Vascular dementia: diagnostic criteria for research studies. Report of the NINDS-AIREN International Workshop. *Neurology* 1993; 43: 250-60.
- Mendez MF. Early-Onset Alzheimer Disease. *Neurol Clin* 2017; 35: 263-81.
- Sommerlad A, Ruegger J, Singh-Manoux A, Lewis G, Livingston G. Marriage and risk of dementia: systematic review and meta-analysis of observational studies. *J Neurol Neurosurg Psychiatry* 2018; 89: 231-8.
- Albanese E, Launer LJ, Egger M, Prince MJ, Giannakopoulos P, Wolters FJ, et al. Body mass index in midlife and dementia: Systematic review and meta-regression analysis of 589,649 men and women followed in longitudinal studies. *Alzheimers Dement (Amst)* 2017; 8: 165-78.
- Wang C, Fu W, Cao S, Jiang H, Guo Y, Xv H, et al. Weight Loss and the Risk of Dementia: A Meta-analysis of Cohort Studies. *Curr Alzheimer Res* 2021; 18: 125-35.
- Cho SH, Jang M, Ju H, Kang MJ, Yun JM, Yun JW. Association of late-life body mass index with the risk of Alzheimer disease: a 10-year nationwide population-based cohort study. *Sci Rep* 2022; 12: 15298.
- Griffiths TD, Lad M, Kumar S, Holmes E, McMurray B, Maguire EA, et al. How Can Hearing Loss Cause Dementia? *Neuron*

- 2020; 108: 401-12.
28. Jang JW, Park JH, Kim S, Lee SH, Lee SH, Kim YJ. Prevalence and Incidence of Dementia in South Korea: A Nationwide Analysis of the National Health Insurance Service Senior Cohort. *J Clin Neurol* 2021; 17: 249-56.
29. Hwangbo S, Lee JY, Han G, Chun MY, Jang H, Seo SW, et al. Dementia incidence and population-attributable fraction for dementia risk factors in Republic of Korea: a 12-year longitudinal follow-up study of a national cohort. *Front Aging Neurosci* 2023; 15: 1126587.
30. Handels R, Jönsson L, Garcia-Ptacek S, Eriksdotter M, Wimo A. Controlling for selective dropout in longitudinal dementia data: Application to the SveDem registry. *Alzheimers Dement* 2020; 16: 789-96.
31. Cox CG, Brush BL, Kobayashi LC, Roberts JS. Determinants of dementia diagnosis in U.S. primary care in the past decade: A scoping review. *J Prev Alzheimers Dis* 2025; 12: 100035.

Evaluating regional diversity in scientific communication: a comparative analysis of COVID-19 preprints and peer-reviewed publications

Dong Hyun Kim^{1*}, Kyulee Jeon^{2,3*}, Seng Chan You^{2,3}

¹Yonsei University College of Medicine, Seoul, Korea

²Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Seoul, Korea

³Yonsei Institute for Digital Health, Yonsei University, Seoul, Korea

Background: The COVID-19 pandemic triggered extensive global research, leading to an unprecedented surge in both peer-reviewed publications and preprints. Despite their widespread use, the implications of preprints for global diversity in scientific communication remain underexplored. This study evaluates how preprints influenced regional diversity in COVID-19 research by analyzing international collaboration networks, social media engagement, and citation patterns compared to peer-reviewed publications.

Methods: We conducted a comparative analysis of COVID-19-related peer-reviewed publications indexed in SCOPUS and preprints from MedRxiv (December 2019–November 2022). Regional diversity was evaluated using bibliometric metrics stratified by World Bank income classifications and geographic regions. International collaboration was quantified using network analysis metrics, while publication, dissemination, and social media engagement were assessed through relative ratios. Social media engagement was measured by quoted posts on X (formerly Twitter). Citation counts were compared between articles with preprint history versus those published directly in journals.

Results: Authors from Sub-Saharan Africa, Latin America, and the Caribbean showed 3.9–4.5 times higher eigenvector centrality in preprints than in peer-reviewed papers, indicating greater integration into global research networks through preprint platforms. Low-income countries showed higher representation in preprints ($P < 0.001$). Although overall engagement on X was similar for both formats, preprints exhibited higher relative ratios of quoted posts across all income groups. Peer-reviewed articles with preprint history received more citations (median=10, 25th–75th percentiles: 3–30) than those without (median=3, 25th–75th percentiles: 0–11, $P < 0.001$), especially from low- and middle-income countries.

Conclusions: Preprints significantly advanced regional diversity in scientific communication during the COVID-19 pandemic. Preprints enhanced international collaboration networks, particularly benefiting researchers from lower-income regions, facilitated broader social media engagement across all income groups, and conferred meaningful citation advantages for subsequent peer-reviewed publications. These results suggest that preprints represent an important mechanism for promoting more equitable participation in global scientific discourse.

Keywords: Preprint; COVID-19; Health equity; Bibliometrics; Social media

Received: April 27, 2025; **Revised:** June 1, 2026; **Accepted:** June 11, 2026

Corresponding author: Seng Chan You

Email: chandryou@yuhs.ac

*These authors contributed equally to this work.

© 2026 Korean Society of Evidence-Based Medicine

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Introduction

The COVID-19 pandemic has triggered extensive research efforts worldwide, with a substantial body of work being produced to understand and combat the virus [1-5]. However, much of this research has been concentrated in high-income countries, leading to disparities in the geographical representation of scientific output. Despite the global nature of the pandemic, researchers from low- and middle-income countries (LMICs) have been underrepresented, further highlighting the existing inequities in the global research landscape.

In this context, preprints have gained attention as a potential solution to overcome the limitations of traditional peer-reviewed journals. The conventional academic publishing process often faces issues such as lengthy and complex review processes, high publication costs, and sometimes bias based on authors' affiliations or nationalities [6]. In contrast, preprints, which involve sharing preliminary manuscripts on public servers before formal peer review, offer advantages such as rapid feedback collection, idea prioritization, and improved access to research [7,8]. The use of preprints surged during the COVID-19 pandemic, particularly in medical research [2,9-11]. The preprint system enables swift dissemination of information in emergency situations, complementing the limitations of traditional publishing systems.

Although numerous preprints have been published via open server during this pandemic [12], little attention has been given to their implications for the inclusive science communication [13]. Meanwhile, social media tools are increasingly part of the research workflow, offering new dissemination and communication possibilities to researchers [14].

The aim of this study is to analyze the impact of COVID-19-related preprints and peer-reviewed papers on global scientific communication. Specifically, we assess the role of preprints in fostering more inclusive research practices by examining regional diversity in authorship, publication, and dissemination. Our analysis draws on data from both publication databases and social media (X, formerly Twitter) to compare the reach and engagement of preprints and peer-reviewed papers, with a focus on regional disparities.

Methods

Data collection

This study collected data from Scopus and medRxiv to assess regional diversity in scientific communication during COVID-19. Scopus is a peer-reviewed literature database, while medRxiv is a health sciences preprint repository [15].

COVID-19-related articles from December 1, 2019, to November 31, 2022, were collected using keywords such as "COVID-19", "SARS-CoV-2", and "coronavirus disease 2019" (Text S1) [1]. By April 20, 2023, metadata including titles, digital object identifiers (DOIs), and author affiliations were obtained, resulting in 277,366 papers from Scopus and 19,584 from medRxiv.

For regional analysis, the author's affiliations were processed using the 'hugofitipaldi/affiliation' R package. X (formerly Twitter) data were integrated using Altmetric and the X/Twitter API, tracking posts linking to papers and categorizing users into demographic groups such as "public", "researcher", "practitioner", and "science communicator". Citation data were retrieved via the CrossRef API to compare citation counts between peer-reviewed papers with and without preprint history [16].

Regional and income-level categorization

For the analysis of regional and income-level diversity, country information was grouped based on the World Bank's 2022 regional and income group classifications [17]. Income groups were defined as follows: low income (gross national income [GNI] per capita of \$1,135 or less), lower-middle income (\$1,136 to \$4,465), upper-middle income (\$4,466 to \$13,845), and high income (\$13,846 or more). Geographical regions were grouped according to the World Bank's administrative classifications, including East Asia and Pacific, Europe and Central Asia, Latin America and the Caribbean, Middle East and North Africa, North America, South Asia, and Sub-Saharan Africa. To address previous findings that Europe and North America are often overrepresented in peer-reviewed journals, we distinguished Europe from Central Asia for clearer analysis.

Global co-authorship networks

To identify patterns of international collaboration between regions, the Inter-regional Co-authorship Count was defined to quantify the number of collaborative papers between researchers from different regions. This was calculated as follows:

$$\text{Inter-regional Co-authorship Count}_{ij} = \sum_{k=1}^N \delta_{ik} \cdot \delta_{jk}$$

where N represents the total number of papers, and δ_{ik} and δ_{jk} indicate whether paper k has authors from regions i and j , respectively.

To evaluate the influence of each region, we used eigenvector centrality (EVC) and calculated the Co-authorship Enhancement Index (CEI), defined as:

$$CEI_r = \frac{EVC_{preprint,r}}{EVC_{peer-reviewed,r}}$$

This ratio indicates whether a region r was more central in the preprint network compared to the peer-reviewed network. Eigenvector centrality assigns scores to nodes (regions) based on their connections to influential nodes, weighted by normalized inter-regional co-authorship counts. This measure highlights which regions hold the most central and influential positions in global research collaboration, allowing us to compare the roles of regions in preprint versus peer-reviewed networks.

Publication of research

To analyze differences in publication between preprints and peer-reviewed articles across different regions, we calculated the Publication Relative Ratio (PubRR) for each country c .

$$PubRR_c = \frac{\left(\frac{\text{Number of preprints from } c}{\text{Total preprints}} \right)}{\left(\frac{\text{Number of peer-reviewed papers from } c}{\text{Total peer-reviewed papers}} \right)}$$

PubRR represents the relative ratio of a country's share of first authors preprints to its share in peer-reviewed papers. This ratio allows us to determine whether researchers in a country are more inclined to publish preprints compared to peer-reviewed articles.

Social media dissemination of research

To evaluate how research is disseminated on social media, we analyzed X posts of preprints and peer-reviewed papers. Social Media Quotation Count quantifies the number of X posts referencing papers from a country. A higher quotation count indicates broader dissemination of that country's research on X.

$$= \sum_{p \in P_c} \text{Social Media Quotation Count}_c$$

Number of X posts citing paper p

where P_c represents the set of papers with first authors from country c .

Social Media Quotation Relative Ratio (QuoRR) compares the share of social media quotation count for preprints versus peer-reviewed articles for each country.

$$QuoRR_c = \frac{\left(\frac{\text{Social media quotations of preprints from } c}{\text{Total social media quotations of all preprints}} \right)}{\left(\frac{\text{Social media quotations of peer-reviewed papers from } c}{\text{Total social media quotations of all peer-reviewed papers}} \right)}$$

Social media readership patterns

To understand how social media users engage with research, we used the Social Media Readership Relative Ratio (ReadRR). ReadRR measures the relative ratio of social media users citing preprints versus peer-reviewed articles from a particular country

$$ReadRR_c = \frac{\left(\frac{\text{Social media quotations from } c \text{ citing preprints}}{\text{Total social media quotations of all preprints}} \right)}{\left(\frac{\text{Social media quotations from } c \text{ citing peer-reviewed papers}}{\text{Total social media quotations of all peer-reviewed papers}} \right)}$$

To explore how the general public engages with research output in each country through social media, we calculated the Public Engagement Relative Ratio (PERR). PERR was used to assess whether public users engage more with preprints than peer-reviewed papers from a given country.

$$PERR_c = \frac{\left(\frac{\text{Public X users' social media quotations citing preprints from } c}{\text{Total social media quotations citing preprints from } c} \right)}{\left(\frac{\text{Public X users' social media quotations citing peer-reviewed papers from } c}{\text{Total social media quotations citing peer-reviewed papers from } c} \right)}$$

Relative ratio interpretation

All relative ratios (PubRR, QuoRR, ReadRR, and PERR) are interpreted as follows:

- *Value > 1: greater engagement for preprints compared to peer-reviewed papers.*
- *Value = 1: equal engagement.*
- *Value < 1: lower engagement for preprints compared to peer-reviewed papers.*

These metrics provide a comprehensive understanding of collaboration, publication, social media dissemination and user engagement across different regions during the COVID-19 pandemic.

Statistical analyses

The coefficient of variation (CV) for inter-regional co-authorship counts was calculated to assess variability in collaboration among regions. Pearson's correlation analyses were performed to examine the relationships between GDP and four relative ratio metrics. Scatter plots and R-squared values were used to visualize these relationships. The Wilcoxon signed-rank test was used to compare the median number of social media posts quoting preprints versus peer-reviewed papers within income groups. A two-sample proportion test was employed to compare the ratio of social media quotation counts to total paper counts across Scopus and medRxiv. Chi-squared tests were used to compare user demographics citing different publication types, and the Wilcoxon rank-sum

test assessed citation differences between peer-reviewed papers with and without preprint versions.

Results

Data collection

We collected 15,413,656 X posts referencing 277,366 papers from Scopus and 2,718,034 X posts referencing 19,584 papers from medRxiv, authored by users from 201 and 213 countries, respectively. For regional diversity analysis, we categorized countries of first authors and X users by income level and geographical region, excluding entries without identifiable country information. The final categorization, which is shown in Fig. S1, included:

1. Country-classified Papers: We categorized papers based on the country of the first author for 250,442 Scopus papers (192 countries) and 13,952 medRxiv papers (147 countries) (Table S1).

2. Country-classified X Posts: We categorized X posts based on the country of the paper's first author (14,325,011 posts for Scopus and 1,853,738 for medRxiv) (Table S2A). Additionally, we categorized posts based on the users' country (5,910,665 posts for Scopus and 1,010,582 for medRxiv) (Table S2B).

Global co-authorship networks

Inter-regional co-authorship counts were calculated across 8 regions, resulting in 28 pairwise combinations for each data source. The average co-authorship count was 196,572.30 for Scopus and 6,833.64 for medRxiv, with a coefficient of variation (CV) of 2.70 for Scopus and 1.95 for medRxiv. This indicates a wider disparity in regional collaboration for peer-reviewed articles compared to preprints. Detailed matrix and counts provided in Fig. S2 and Table S3. The co-authorship enhancement index (CEI)—calculated as the ratio of EVC for preprints to peer-reviewed publications—was highest in North America (7.66), followed by Sub-Saharan Africa (4.53), Central Asia (4.29), Latin America & Caribbean (3.90), Middle East & North Africa (2.19), East Asia & Pacific (1.76), and South Asia (1.07), with Europe serving as the baseline (EVC = 1.00) (Table 1).

Publication of research

The Publication Relative Ratio (PubRR) was calculated to compare the number of papers authored by researchers from each country across Scopus and medRxiv. A significant negative correlation between PubRR and GDP was observed (Pearson correlation coefficient: -0.38, P value < 0.001), indicating that researchers from countries with lower GDPs are more likely to publish in preprint repositories (Fig. 1A).

Over half of the countries in North America, Latin America, Central Asia, and Sub-Saharan Africa displayed PubRR values greater than 1 (Fig. 1B, Fig. S3). High PubRR values were recorded in countries such as Turkmenistan (35.9), South Sudan (33.3), Mali (15.9), Cameroon (11.2) and Ukraine (9.9) (Table S4).

Social media dissemination of research

We analyzed the distribution of posts on platform X that referenced preprints and peer-reviewed papers by first authors from each country. Across all income groups, the median Social Media Quotation Count for medRxiv was significantly higher than for Scopus (Wilcoxon test, P value < 0.001) (Fig. 2A, Table S5). A significant negative correlation was found between Social Media Quotation Relative Ratio (QuoRR) and GDP (Pearson correlation coefficient: -0.17, P value = 0.046), indicating that countries with lower GDPs had a higher relative share of social media quotations for preprints (Fig. 2B). Approximately half of the countries in North America, Sub-Saharan Africa, and Middle East & North Africa had QuoRR values greater than 1, suggesting relatively higher social media attention for preprints in these regions (Fig. 2C, Fig. S5). High QuoRR values were observed in countries like Cameroon (80.8), South Sudan (62.9), Sudan (36.7), Turkmenistan (19.3), and Kazakhstan (13.1) (Table S6).

Table 1. Regional Comparison of International Collaboration Patterns in Preprints Versus Peer-Reviewed Publications

	Peer-reviewed papers (EVC)	Preprints (EVC)	Co-authorship Enhancement Index
Europe	1.000	1.000	1.00
North America	0.106	0.812	7.66
East Asia & Pacific	0.077	0.135	1.76
Latin America & Caribbean	0.026	0.103	3.90
Middle East & North Africa	0.040	0.088	2.19
Sub-Saharan Africa	0.012	0.055	4.53
South Asia	0.031	0.033	1.07
Central Asia	0.008	0.003	4.29

Eigenvector centrality (EVC) values reflect each region's influence in international collaboration networks, weighted by connections to scientifically significant regions. The Collaboration Enhancement Index (CEI) is calculated by dividing preprint EVC by peer-reviewed publication EVC, quantifying the relative increase in international collaboration intensity through preprints. Regions with the most substantial collaboration enhancement through preprints include North America (7.66), Sub-Saharan Africa (4.53), Central Asia (4.29), Latin America & the Caribbean (3.9), and the Middle East & North Africa (2.19).

EVC, Eigenvector Centrality; CEI, Co-authorship Enhancement Index.

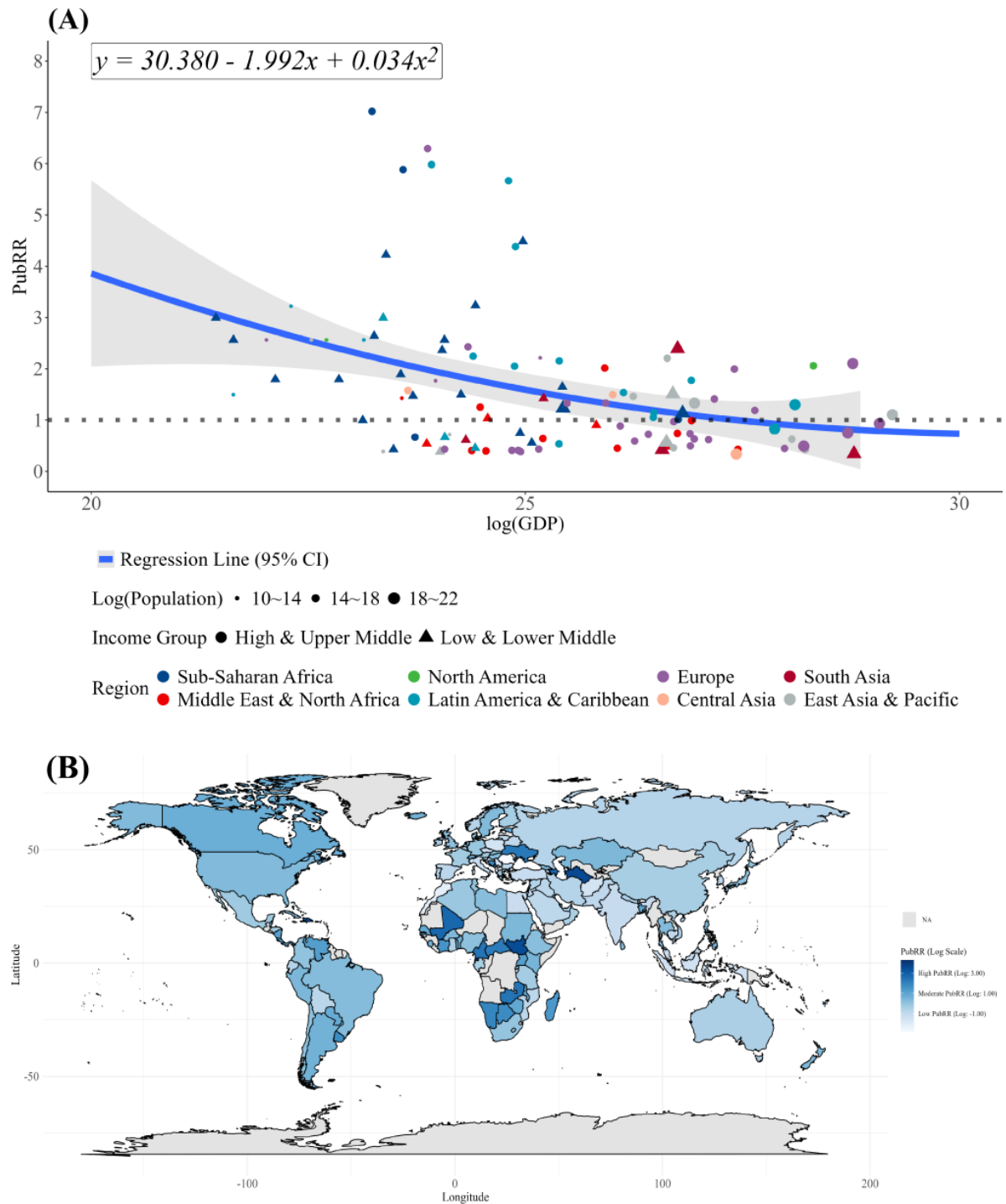


Fig. 1. Global distribution of relative preprint publication ratios by economic indicators and geographic regions (A) Scatter plot of country GDP and Publication Relative Ratio (PubRR), (B) Geographic heatmap of PubRR values worldwide. The figure shows the relationship between countries' economic indicators and their relative preprint publication ratios. PubRR was calculated based on the count of papers published in each country according to the first author's affiliation, with top and bottom 10% outliers excluded. In panel A, the scatter plot displays log(GDP) on the x-axis and PubRR values on the y-axis, with point size representing log(Population) and colors indicating income classification (HMIC/LMIC) and geographic region. The second-order polynomial trend line shows a negative correlation (Pearson correlation coefficient -0.28, P value < 0.001). Panel B presents PubRR values on a geographic heatmap, with countries lacking PubRR data shown in gray. GDP, Gross Domestic Product; PubRR, Publication Relative Ratio; CI, Confidence Interval; HMIC, High-Middle Income Countries; LMIC, Low-Middle Income Countries; NA, not available.

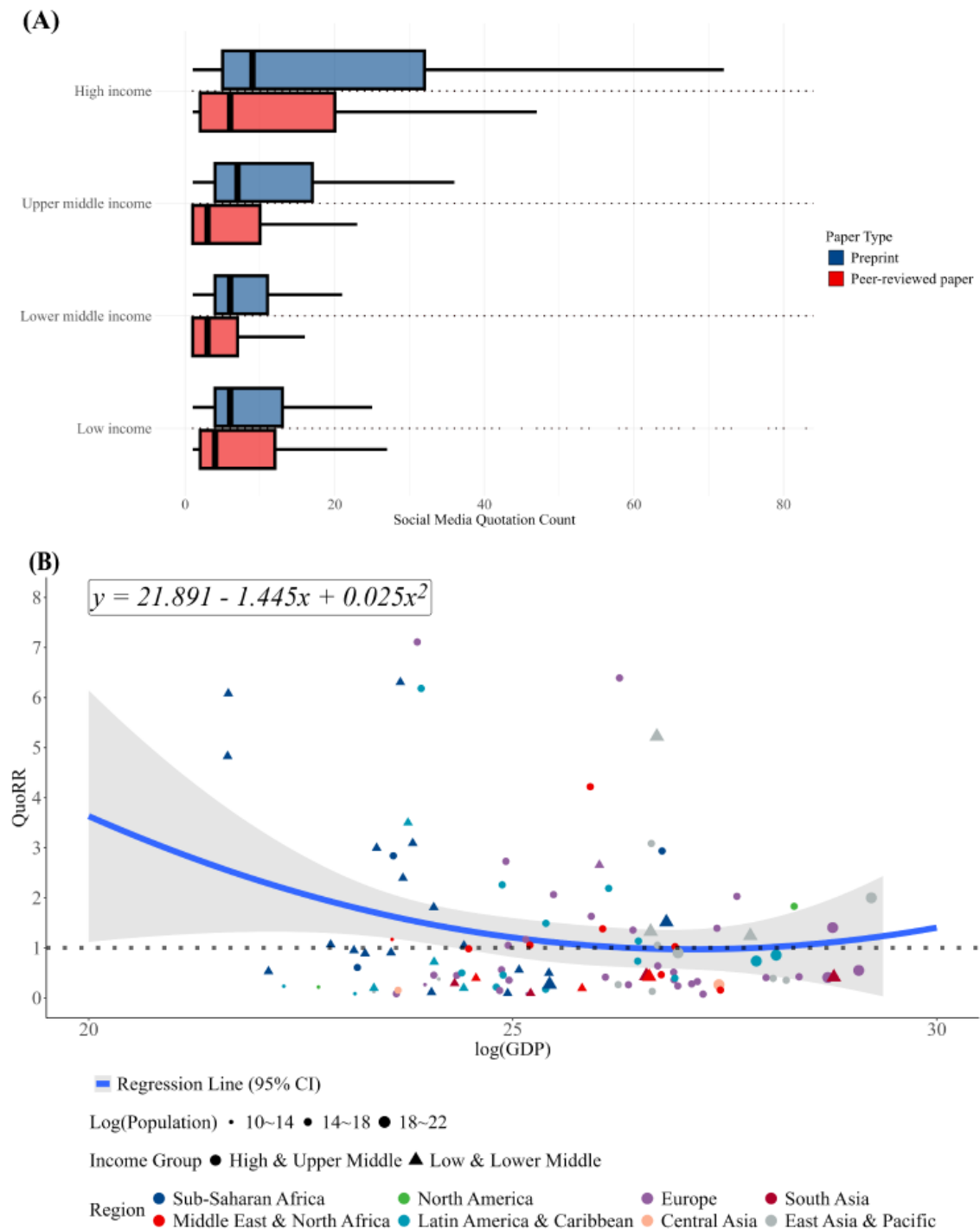


Fig. 2. Social media engagement patterns for COVID-19 publications by country income level and geographic distribution (A) Boxplot of Social Media Quotation Count by Income and Paper Type, (B) Scatter plot of the country's GDP and Social Media Quotation Relative Ratio (QuoRR), (C) Geographical Heatmap of QuoRR

Fig. 2 (A) is a set of boxplots for each paper type (preprints and peer-reviewed papers) based on the social media quotation counts of each paper. They are sub-grouped by the income group of the 1st author's country. Outliers were excluded when drawing these boxplots. Five-number summaries for each boxplot starting from the top are (1, 5, 9, 32, 72), (1, 2, 6, 20, 47), (1, 4, 7, 17, 36), (1, 1, 3, 10, 23), (1, 4, 6, 11, 21), (1, 1, 3, 7, 16), (1, 4, 6, 13, 25) and (1, 2, 4, 12, 27). Fig. 2 (B), (C) was made by examining the social media quotation counts of papers published in each country then calculating QuoRR.

GDP, Gross Domestic Product; QuoRR, Social Media Quotation Relative Ratio; CI, Confidence Interval; Log, Common Logarithm; HMIC, High-Middle Income Countries; LMIC, Low-Middle Income Countries; NA, not available. (Continued to the next page)

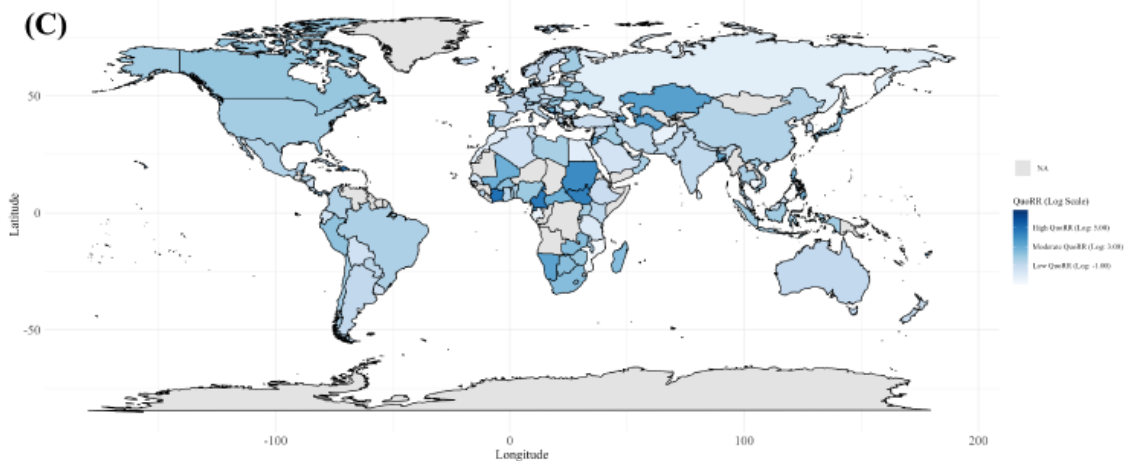


Fig. 2. Continued.

Social media readership patterns

We compared social media engagement for preprints versus peer-reviewed papers across countries. No significant correlation was found between GDP and the Social Media Readership Relative Ratio (ReadRR) (Pearson correlation coefficient: -0.11, P value = 0.136) (Fig. S6, Fig. S7). Both paper types exhibit a similar decline in the social media quotation-to-paper ratio from high- to low-income groups, with no significant difference observed (Fig. S8). The general public was the most frequent group citing both preprints and peer-reviewed papers, with a higher engagement rate for preprints (preprints: 88.26%, peer-reviewed papers: 85.75%, P value < 0.001) (Fig. S9). Scientists, practitioners and science communicators showed greater engagement with peer-reviewed articles (P value < 0.001). The Public Engagement Relative Ratio (PERR) exhibited a negative correlation with GDP (Pearson correlation coefficient: -0.20, P value = 0.021). In North America and Central Asia, all countries had PERR values greater than 1. Similarly, in East Asia & Pacific, Sub-Saharan Africa, and Latin America & Caribbean, approximately 90% of countries had PERR greater than 1 (Fig. S10). High PERR values were observed in countries like Algeria (1.2), Palestine (4.4), Cameroon (1.3), Kazakhstan (1.2), Armenia (1.2), and Cambodia (1.2) (Table S7).

Citation analyses

We assessed the impact of prior preprint publication on subsequent citation counts in peer-reviewed journals. Papers with a preprint history had significantly higher citation counts (median=10 [Q1-Q3: 3-30]) compared to those without (median=3 [Q1-Q3: 0-11]) (P value < 0.001) (Fig. 3). This trend held true for papers from low- and middle-income countries (LMICs), with preprints receiving higher citation

counts (median=8 [Q1-Q3: 2-22]) than those without (median=2 [Q1-Q3: 0-8]) (P value < 0.001). The median citation counts by country for preprints and peer-reviewed papers are shown in Fig. S11 and citation count statistics by region and income group for peer-reviewed papers and preprints in Table S8. We also tracked changes in the income classification of first authors when preprints transitioned to peer-reviewed publications. Of these, 57% of preprints from low-income countries, 49% from lower-middle-income countries, and 22% from upper-middle-income countries were subsequently published with first authors from higher-income groups. Among LMIC preprints, 22 remained published by LMIC authors, while 26 transitioned to authorship from high-income countries (Table S9).

Discussion

This study provides a comprehensive bibliometric analysis of how preprints influenced regional diversity in scientific communication during the COVID-19 pandemic. Our comparative analysis of preprints and peer-reviewed publications revealed distinct patterns across geographical regions and economic strata. Three key findings emerged: First, preprints significantly enhanced international collaboration by facilitating the inclusion of authors from regions traditionally underrepresented in peer-reviewed publications. Second, researchers from lower-GDP countries showed higher engagement with preprint platforms, suggesting these platforms serve as a more accessible publication route. Third, preprints demonstrated broader reach through social media engagement and, notably, papers that began as preprints achieved higher citation counts after peer review, particularly those from lower-GDP regions. These findings collectively indicate

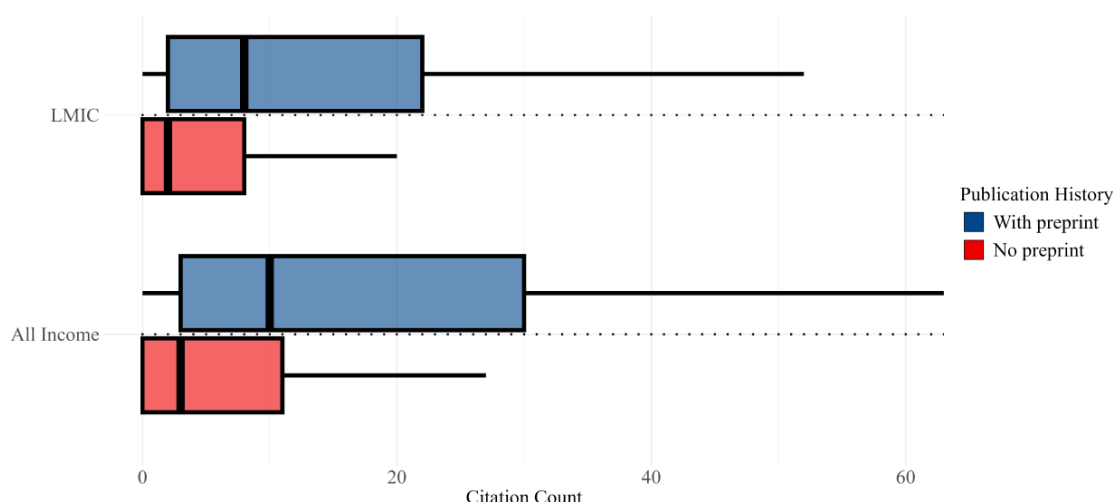


Fig. 3. Boxplot of citation count by preprint publication history

The figure shows citation count distribution of peer-reviewed papers with and without preprint history. Peer-reviewed papers with preprint history exhibited higher citation counts (median=10, Q1–Q3: 3–30) compared to those without (median=3, Q1–Q3: 0–11), P value < 0.001. This pattern was also observed in LMIC countries, where papers with preprint history received more citations (median=8, Q1–Q3: 2–22) than those without (median=2, Q1–Q3: 0–11), P value < 0.001.

LMIC, Low-Middle Income Countries.

that preprints serve as an effective mechanism for democratizing scientific communication, enabling researchers from underrepresented regions to increase their visibility and impact in the global scientific discourse.

Our study found that peer-reviewed publications tend to rely heavily on established collaborative networks, especially in affluent regions such as North America, Europe, and East Asia. This reliance may be attributed to long-standing research relationships [18–20], greater access to funding [21,22], and the prestige of established institutions [23–25]. In contrast, our analysis revealed that preprints demonstrated a more balanced distribution of collaborations across regions, with particularly strong representation from lower-GDP countries. The higher utilization of preprint servers by researchers from lower-GDP countries compared to traditional peer-reviewed journals suggests that preprints provide a more accessible publication route for these researchers.

This accessibility proved especially crucial during the COVID-19 pandemic [26], where rapid dissemination of research findings was essential for global public health responses [27]. By enabling researchers from lower-income regions to contribute more readily to the scientific discourse, preprints helped foster greater inclusivity in scientific communication during a critical period.

Our analysis further revealed that preprints were disseminated more extensively on social media, particularly by public users. This trend was especially pronounced among papers by researchers from lower-GDP countries, for whom

social media served as a crucial tool to enhance visibility and facilitate engagement with a broader audience [28,29]. The role of social media in disseminating research is critical, as it provides an alternative avenue for researchers who may lack the traditional means of publication and collaboration [28,30]. Previous studies have shown that social media can significantly increase the citation impact of research, particularly for authors from institutions with fewer resources [31,32]. This suggests that the use of social media not only facilitates the dissemination of research but also democratizes access to scientific knowledge, allowing researchers from lower GDP countries to engage more actively in global discourse.

Our analyses also indicate that peer-reviewed papers that originated as preprints tend to receive a significantly higher median citation count compared to those that did not. This trend is especially pronounced for papers where LMIC researchers are the first authors. The increased visibility and dissemination afforded by preprints may enhance subsequent citations, as preprints often serve as a preliminary platform for sharing research findings, which can lead to greater engagement and discourse within the academic community [33–35]. However, it has been observed that LMIC authorship in preprints often shifts to higher-income countries in peer-reviewed versions, likely due to disparities in funding, institutional support, and perceived credibility [21,36]. This raises concerns about equity and representation in global research outputs [37,38].

Despite the advantages of preprints, concerns regarding their quality and credibility persist due to the absence of formal peer review prior to publication. Addressing these concerns requires fostering an open discussion environment with interactive feedback mechanisms [39]. Utilizing communication tools such as forums and collaborative platforms can enhance the quality assurance of preprints by allowing researchers to engage in constructive dialogue about the findings presented [40]. This approach is supported by the notion that preprints can be revised based on community input. Furthermore, the integration of communication tools such as social media can facilitate these discussions, allowing researchers to share insights and critiques in real-time. The use of interactive feedback mechanisms can help establish frameworks, allowing for a more robust evaluation of preprints. This aligns with the findings of Weissgerber et al., who noted that the overwhelming influx of preprints during the pandemic necessitated better monitoring and assessment strategies [41]. By leveraging communication tools, researchers can create a collaborative environment that not only addresses quality concerns but also fosters a culture of transparency and accountability in scientific communication.

Although this study provides valuable insights, there are some limitations. One limitation relates to dataset attrition bias. A total of 9.7% of peer-reviewed papers and 28.8% of preprints were excluded due to incomplete affiliation details, despite the use of advanced methods to extract country information. This higher exclusion rate for preprints is largely due to the preprint system's reliance on free-format entries for affiliations, lacking the standardized format used in peer-reviewed papers, which complicates accurate geographic extraction. Additionally, attempts to automate web-API searches were hindered by inconsistent author identification. The reliance on SCOPUS for peer-reviewed papers and medRxiv for preprints may have excluded relevant publications from other platforms [42]. English-language queries may have missed non-English titles, particularly early COVID-19 studies from China [43]. Furthermore, the country data were based on first authors' affiliations rather than their origin, and this study did not analyze time trends or journal impact factors, which could further refine the findings. Future research could address these aspects.

Conclusions

The COVID-19 pandemic has clearly demonstrated how inequalities in scholarly communication can be exacerbated during a global health crisis. While it is unlikely that preprints will fully replace peer-reviewed journals in the medical field,

they have proven to be an important tool in mitigating these inequalities. With specific and measurable objectives, collaborative efforts can contribute to the development of an inclusive and sustainable scientific communication environment.

Conflict of Interest

SCY reports being a chief executive officer of PHI Digital Healthcare; and grants from Daiichi Sankyo. He is a coinventor of granted Korea Patent DP-2023-1223 and DP-2023-0920, and pending Patent Applications DP-2024-0909, DP-2024-0908, DP-2022-1658, DP-2022-1478, and DP-2022-1365, unrelated to current work.

Funding

This study was supported by the NAVER Digital Bio Innovation Research Fund, funded by NAVER Corporation (Grant No. [3720230020]).

Data Availability Statement

The supplementary files, code for web-scraping medRxiv and Altmetric data, preprocessing scripts, and ratio calculation methodology are available at https://github.com/dr-you-group/SPARK_COVID.git. The repository also includes raw text data of X (formerly Twitter) preprint quotations collected for analysis. Additional details regarding data collection and analytical methods are available upon reasonable request from the corresponding author.

Ethics Approval and Consent to Participate

This study analyzed publicly available bibliometric data from SCOPUS and medRxiv, and social media data from X (formerly Twitter). No human subjects were directly involved in this research, and no personal identifiable information was collected or analyzed. Therefore, this study was exempt from institutional review board approval. The conduct and reporting of this research were guided by the principles set forth by the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines for cross-sectional studies.

Authors' Contributions

Conceptualization: DHK KLJ SCY, Data curation: DHK, Formal analysis: DHK KLJ, Investigation: DHK KLJ, Methodology: DHK KLJ SCY, Project administration: SCY, Resources:

SCY, Software: DHK, Supervision: SCY, Validation: DHK KLJ
 SCY, Visualization: DHK KLJ, Writing – original draft: DHK
 KLJ, Writing – review & editing: DHK KLJ SCY

Acknowledgments

We extend our sincere gratitude to Dr. Joshua Wallach from Emory University for his invaluable insights and constructive feedback that significantly enhanced this study.

Supplementary Materials

Supplementary files are available from <https://doi.org/10.63528/jebp.2026.00007>

ORCID

Dong Hyun Kim, <https://orcid.org/0009-0002-5618-1472>

Kyulee Jeon, <https://orcid.org/0009-0008-6338-5522>

Seng Chan You, <https://orcid.org/0000-0002-5052-6399>

References

- Ioannidis JP, Salholz-Hillel M, Boyack KW, Baas J. The rapid, massive growth of COVID-19 authors in the scientific literature. *R Soc Open Sci* 2021; 8: 210389.
- Else H. How a torrent of COVID science changed research publishing - in seven charts. *Nature* 2020; 588: 553.
- Majumder MS, Mandl KD. Early in the epidemic: impact of preprints on global discourse about COVID-19 transmissibility. *Lancet Glob Health* 2020; 8: e627-8.
- Stoye E. China coronavirus: how many papers have been published? *Nature* [serial on the Internet]. 2020 Feb 5 Available from <https://www.nature.com/articles/d41586-020-00253-8>
- Kana MA, LaPorte R, Jaye A. Africa's contribution to the science of the COVID-19/SARS-CoV-2 pandemic. *BMJ Glob Health* 2021; 6: e004059.
- Krumholz HM. The end of journals. *Circ Cardiovasc Qual Outcomes* 2015; 8: 533-4.
- Bagdasarian N, Cross GB, Fisher D. Rapid publications risk the integrity of science in the era of COVID-19. *BMC Med* 2020; 18: 192.
- Ross GJ, Bloom T. Preprints can fill a void in times of rapidly changing science. *STAT* [serial on the Internet]. 2020 Jan 31 [cited 2026 Sep 3]. Available from <https://www.statnews.com/2020/01/31/preprints-fill-void-rapidly-changing-science/>
- Flanagin A, Fontanarosa PB, Bauchner H. Preprints involving medical research—do the benefits outweigh the challenges?. *JAMA* 2020; 324: 1840-3.
- Fraser N, Brierley L, Dey G, Polka JK, Palfy M, Nanni F, et al. Preprinting the COVID-19 pandemic. *PLoS Biol* 2021; 19: e3000959.
- Rubin EJ, Baden LR, Morrissey S, Campion EW. Medical journals and the 2019-nCoV outbreak. *N Engl J Med* 2020; 382: 866.
- Iyandemye J, Thomas MP. Low income countries have the highest percentages of open access publication: a systematic computational analysis of the biomedical literature. *PLoS One* 2019; 14: e0220229.
- Priem J, Costello KL. How and why scholars cite on Twitter. *Proc Am Soc Inf Sci Technol* 2010; 47: 1-4.
- Ferreira MR, Mongeon P, Costas R. Large-scale comparison of authorship, citations, and tweets of Web of Science authors. *J Altmetrics* 2021; 4: 1.
- Falagas ME, Pitsouni EI, Malietzis GA, Pappas G. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. *FASEB J* 2008; 22: 338-42.
- Crossref. REST API [Internet]. Lynnfield (MA): Crossref; [cited 2026 Sep 3]. Available from <https://www.crossref.org/documentation/retrieve-metadata/rest-api/>
- World Bank. World Bank country and lending groups [Internet]. Washington (DC): World Bank; [cited 2026 Sep 3]. Available from <https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups>
- Bu Y, Ding Y, Liang X, Murray DS. Understanding persistent scientific collaboration. *J Assoc Inf Sci Technol* 2018; 69: 438-48.
- Chinchilla-Rodriguez Z, Ferligoj A, Miguel S, Kronegger L, de Moya-Anegón F. Blockmodeling of co-authorship networks in library and information science in Argentina: a case study. *Scientometrics* 2012; 93: 699-717.
- Guerrero Bote VP, Olmeda-Gomez C, De Moya-Anegón F. Quantifying the benefits of international scientific collaboration. *J Am Soc Inf Sci Technol* 2013; 64: 392-404.
- Smith MJ, Weinberger C, Bruna EM, Allesina S. The scientific impact of nations: journal placement and citation performance. *PLoS One* 2014; 9: e109195.
- Tubbs-Cooley HL, Martsolf DS, Pickler RH, Morrison CF, Wardlaw CE. Development of a regional nursing research partnership for academic and practice collaborations. *Nurs Res Pract* 2013; 2013: 473864.
- Chinchilla-Rodriguez Z, Sugimoto CR, Larivière V. Follow the leader: on the relationship between leadership and scholarly impact in international collaborations. *PLoS One* 2019; 14: e0218309.
- Onyancha OB. Regional and international research collaboration and citation impact in selected sub-Saharan African countries in the period 2000 to 2019. *Glob Knowl Mem Commun*

- 2020; 70: 577-94.
25. Limaymanta CH, Quiroz-de-Garcia R, Rivas-Villena JA, Rojas-Arroyo A, Gregorio-Chaviano O. Relationship between collaboration and normalized scientific impact in South American public universities. *Scientometrics* 2022; 127: 6391-411.
 26. Gianola S, Jesus TS, Barger S, Castellini G. Characteristics of academic publications, preprints, and registered clinical trials on the COVID-19 pandemic. *PLoS One* 2020; 15: e0240123.
 27. Teixeira da Silva JA, Bornemann-Cimenti H, Tsigaris P. Optimizing peer review to minimize the risk of retracting COVID-19-related literature. *Med Health Care Philos* 2021; 24: 21-6.
 28. Giustini DM, Ali SM, Fraser M, Kamel Boulos MN. Effective uses of social media in public health and medicine: a systematic review of systematic reviews. *Online J Public Health Inform* 2018; 10: e62121.
 29. Gates A, Featherstone R, Shave K, Scott SD, Hartling L. Dissemination of evidence in paediatric emergency medicine: a quantitative descriptive evaluation of a 16-week social media promotion. *BMJ Open* 2018; 8: e022298.
 30. Tunnecliff J, Ilic D, Morgan P, Keating J, Gaida JE, Clearihan L, et al. The acceptability among health researchers and clinicians of social media to translate research evidence to clinical practice: mixed-methods survey and interview study. *J Med Internet Res* 2015; 17: e119.
 31. El Rassi R, Meho LI, Nahlawi A, Salameh JS, Bazarbachi A, Akl EA. Medical research productivity in the Arab countries: 2007-2016 bibliometric analysis. *J Glob Health* 2018; 8: 020411.
 32. Dol J, Tutelman PR, Chambers CT, Barwick M, Drake EK, Parker JA, et al. Health researchers' use of social media: scoping review. *J Med Internet Res* 2019; 21: e13687.
 33. Bertin M, Atanassova I. Preprint citation practice in PLOS. *Scientometrics* 2022; 127: 6895-912.
 34. Traag VA. Inferring the causal effect of journals on citations. *Quant Sci Stud* 2021; 2: 496-504.
 35. Subramanian K, Nalli A, Senthil V, Bhat A. Pharmaceutical industry-authored preprints: scientific and social media impact. *Curr Med Res Opin* 2021; 37: 269-73.
 36. Eldridge L, Garton EM, Duncan K, Gopal S. Authorship of publications supported by NCI-funded grants involving low- and middle-income countries. *JAMA Netw Open* 2024; 7: e243215.
 37. Sofi-Mahmudi A, Stojanova J, Vounzoulaki E, Tomlinson E, Beatriz-Pizarro A, Khademioore S, et al. Geographical and gender diversity in Cochrane and non-Cochrane reviews authorship: a meta-research study. *medRxiv [Preprint]* 2024; Mar 25 [cited 2026 Sep 3]. Available from <https://www.medrxiv.org/content/10.1101/2024.03.23.24304672v2>.
 38. Schneider H, Maleka N. Patterns of authorship on community health workers in low- and middle-income countries: an analysis of publications (2012-2016). *BMJ Glob Health* 2018; 3: e000797.
 39. Blatch-Jones A, Saucedo AR, Giddins B. The use and acceptability of preprints in health and social care settings: a scoping review. *PLoS One* 2023; 18: e0291627.
 40. Zeng L. Changes in health communication in the age of COVID-19: a study on the dissemination of preprints to the public. *Front Public Health* 2023; 11: 1078115.
 41. Weissgerber T, Riedel N, Kilicoglu H, Labbe C, Eckmann P, Ter Riet G, et al. Automated screening of COVID-19 preprints: can we help authors to improve transparency and reproducibility? *Nat Med* 2021; 27: 6-7.
 42. Irawan DE, Zahroh H, Puebla I. Preprints as a driver of open science: opportunities for Southeast Asia. *Front Res Metr Anal* 2022; 7: 992942.
 43. Fidahic M, Nujic D, Runjic R, Civljak M, Markotic F, Lovric Makaric Z, et al. Research methodology and characteristics of journal articles with original data, preprint articles and registered clinical trial protocols about COVID-19. *BMC Med Res Methodol* 2020; 20: 161.

Current status and challenges of Evidence-Based Medicine education in Korean medical schools: a cross-sectional survey of 21 institutions

Hyun Jung Kim^{1,2}, Mookyung Oh³, Jae Hung Jung⁴, Geun Joo Choi⁵, Heather Swan^{1,2}, Soo Young Kim⁶

¹Cochrane Korea, Seoul, Korea

²Institute for Evidence-Based Medicine, College of Medicine, Korea University, Seoul, Korea

³Department of Internal Medicine, Soonchunhyang University Cheonan Hospital, Cheonan, Korea

⁴Department of Precision Medicine and Urology, Yonsei University Wonju College of Medicine, Wonju, Korea

⁵Department of Anesthesiology and Pain Medicine, Chung-Ang University College of Medicine, Seoul, Korea

⁶Department of Family Medicine, Kangdong Sacred Heart Hospital, Hallym University College of Medicine, Seoul, Korea

Background: Evidence-based medicine (EBM) is an important element of medical education. However, nationwide data on EBM education in Korean medical schools are limited. This study aimed to describe the organizational structure, content, methods, and assessment of EBM education, together with barriers and support needs, and to explore factors associated with variation across schools.

Methods: A 26-item questionnaire was administered to EBM education leaders or course directors at 21 Korean medical schools. Data were analyzed descriptively, and differences by structural and exogenous characteristics were explored using cross-tabulation.

Results: Twenty-one of 40 medical schools responded. All responding institutions offered EBM as a mandatory subject, but 42.9% had no dedicated organizing unit. Coverage of classic EBM steps was high (formulating clinical questions: 95.2%; literature searching: 95.2%; critical appraisal: 76.2%), whereas AI-assisted evidence summarization was included in only 19.0%. Only 19.0% reported systematic theory-clinical integration, and 57.1% left bedside EBM to preceptor discretion. Demand for a standardized curriculum guide (90.5%) and faculty development (76.2%) was high, and 100% were willing to use externally developed materials. Variation in key outcomes was associated with organizational governance but was not explained by exogenous structural characteristics such as class size, clinical infrastructure, establishment type, or region.

Conclusions: EBM education in Korean medical schools is universally mandatory but uneven in organization, theory-clinical integration, faculty capacity, assessment, and AI integration. These differences appeared to reflect governance more than resource size, suggesting that society-level support should include organizational models alongside standardized curricula, assessment tools, and faculty development.

Keywords: Evidence-based medicine; Medical education; Medical Schools; Curriculum; Surveys and questionnaires; Republic of Korea

Introduction

Evidence-based medicine (EBM) is an approach to clinical decision-making that integrates the best research evidence

with clinical expertise and patient values, and since the 1990s it has become a core competency in medical education [1,2]. EBM education is typically structured around five steps: formulating clinical questions (PICO), acquiring evidence

Received: April 27, 2025; **Revised:** June 6, 2026; **Accepted:** June 12, 2026

Corresponding author: Soo Young Kim

Email: hallymfm@gmail.com

© 2026 Korean Society of Evidence-Based Medicine

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

(literature searching), critical appraisal, application, and evaluation of performance. There is international consensus that it should not be confined to lectures but linked to clinical settings and reinforced longitudinally across the curriculum [3].

Korean medical education is undergoing a major transition. The February 2024 amendment to the Enforcement Decree of the Higher Education Act removed the distinction between a 2-year premedical and a 4-year medical phase, allowing each school to autonomously design an integrated 6-year curriculum [4]. The accreditation standards of the Korean Institute of Medical Education and Evaluation, ASK2026, explicitly require evidence-based reasoning and decision-making competency (K.2.2.1), basic-clinical vertical integration (K.2.3.2), and modular/spiral design (K.2.1.1) [5]. Within this context, it has been argued that EBM should be positioned not as knowledge confined to a single course but as a longitudinal competency running through the 6-year curriculum [6,7].

Despite these institutional demands, systematic nationwide data on how Korean medical schools actually organize, deliver, and assess EBM are scarce. Although individual curricular reform cases have been reported [8,9], surveys encompassing the responsible body, timing and hours, inclusion of each step, theory-clinical integration, assessment, and the barriers and support needs perceived in the field are rare.

Therefore, this study surveyed Korean medical schools to describe (1) the organizational structure and mode of delivery, (2) content and methods by step, (3) theory-clinical integration and assessment, (4) barriers and support needs of EBM education, and (5) to explore which structural factors and characteristics were associated with variation in outcomes, in order to provide baseline data for future standardization and curricular integration.

Methods

Study Design and Participants

This was a cross-sectional survey to assess the status of EBM education in Korean medical schools. With the cooperation of the Korean Society for Evidence-Based Medicine and the Korean Society of Medical Education, the survey was distributed to all 40 medical schools nationwide and administered via Google Forms.

Survey Instrument

The questionnaire comprised 26 items covering (1) institutional and respondent characteristics; (2) the body governing

EBM education and the mode, timing, and hours of delivery; (3) inclusion of content for each EBM step (formulating clinical questions, literature searching, critical appraisal, clinical practice guidelines, levels of evidence/strength of recommendation, and AI-based evidence summarization); (4) teaching methods and use of digital/AI tools; (5) theory-clinical integration and involvement of specialists; (6) assessment methods and reflection in graduation requirements; (7) operational difficulties and barriers; and (8) demand for society-level support materials and willingness to use external materials. Items used a mix of single-choice, multiple-choice, and free-text formats. To develop the instrument, a preliminary status review was conducted with faculty members responsible for EBM education at several medical schools, and the content validity of the items was reviewed by the board of directors of the Korean Society for Evidence-Based Medicine.

Data Collection and Analysis

Data were collected via an online survey between April 1 and April 30, 2026. To improve the response rate, an official request was made through the Korean Society of Medical Education. Given the small sample size, analyses were primarily descriptive (frequencies and percentages), and 95% confidence intervals for key proportions were calculated using the Wilson method. For multiple-choice items, the proportion selecting each option was calculated using the number of responding institutions ($n=21$) as the denominator. To explore factors associated with variation in outcomes, EBM governance was operationalized as “having a dedicated organizing unit” and “an integrated methodology-clinical design,” and was cross-tabulated with outcome indicators and exogenous structural characteristics (class size, establishment type, region, and number of tertiary hospitals). For this, the 2026 admission quota (based on the reversion to the pre-expansion level of 3,058) and the Ministry of Health and Welfare’s 5th-cycle (2024–2026) designation of tertiary hospitals were used as external data, and a sensitivity analysis weighted by admission quota was performed in parallel. Fisher’s exact test was applied to selected bivariate relationships; however, because no correction for multiple comparisons was applied and statistical power was limited, p -values were interpreted only for hypothesis generation rather than confirmation. Owing to constraints on sample and event numbers, multivariable analysis was not performed. Free-text responses were summarized qualitatively.

Results

Organizational Structure of EBM Education

The respondents were one person per institution responsible for or involved in EBM education at each of the 21 schools. Eight respondents (38.1%) were medical education leaders (e.g., associate dean for education), 7 (33.3%) were course directors (e.g., EBM/clinical epidemiology), 5 (23.8%) were 'other', and one (4.8%) was a clinical clerkship director (Table 1). As of the 2026 academic year, there are 40 medical schools in Korea (admission quota 3,058); the 21 responding schools represent 52.5% of institutions and 56.8% (1,738/3,058) of the admission quota. Response rates were relatively higher among national universities (7/10) and schools in the Seoul metropolitan area (9/13). All responding institutions (21/21, 100%) provided EBM as mandatory education (Table 2). However, the most common arrangement was having no dedicated organizing unit (9 schools, 42.9%), followed by a specific department (e.g., preventive medicine, health policy and management, biostatistics) in 7 (33.3%), an associate-dean-led structure in 4 (19.0%), and a department of medical education in 1 (4.8%) (Table 1). The most common mode of delivery was as part of an integrated course (10

schools, 47.6%), followed by integrated methodology-clinical teaching (7, 33.3%) and clinical-department-led teaching (3, 14.3%). EBM was most often taught in medical years 1–2 (57.1%) and years 3–4 (38.1%), with early exposure in the pre-medical phase in only 14.3%. Total instructional hours were ≥ 21 hours and 11–20 hours in 33.3% each, whereas ≤ 10 hours also accounted for 33.3%, indicating substantial variation across institutions.

Educational Content

Coverage by EBM step was highest for formulating clinical questions (PICO, 95.2%) and literature searching (95.2%), followed by interpreting levels of evidence/strength of recommendation (81.0%), critical appraisal of research (76.2%), and appraisal/use of clinical practice guidelines (61.9%), indicating generally high coverage of the classic EBM steps (Table 3). In contrast, only 19.0% included education on

Table 1. Characteristics of Responding Institutions and Respondents (n=21)

Characteristic	Category	n (%)
No. of responding institutions		21
Respondent role	Medical education leader (e.g., associate dean)	8 (38.1)
	Course director (EBM/clinical epidemiology, etc.)	7 (33.3)
	Other	5 (23.8)
	Clinical clerkship director	1 (4.8)
Body governing EBM	No dedicated unit	9 (42.9)
	Specific department (preventive medicine, health policy/management, biostatistics, etc.)	7 (33.3)
	Associate-dean-led	4 (19.0)
	Department of medical education	1 (4.8)

Table 2. Organizational Structure of EBM Education (n=21)

Item	Category	n (%)
Mode of delivery (Q4)	Part of an integrated course	10 (47.6)
	Integrated methodology-clinical teaching	7 (33.3)
	Clinical-department-led teaching	3 (14.3)
	Other	1 (4.8)
Delivery format (Q5, multiple)	Part of an integrated course	14 (66.7)
	During clinical clerkship	11 (52.4)
	Within research-methods/statistics course	11 (52.4)
	Standalone course	6 (28.6)
Timing (Q6, multiple)	Medical years 1–2	12 (57.1)
	Medical years 3–4	8 (38.1)
	Across all years	7 (33.3)
	Clinical clerkship	4 (19.0)
	Premedical	3 (14.3)
Total hours (Q7)	≥ 21 hours	7 (33.3)
	11–20 hours	7 (33.3)
	1–5 hours	4 (19.0)
	6–10 hours	3 (14.3)
Mandatory (Q8)	Mandatory	21 (100)

Table 3. Inclusion of Content by EBM Step (n=21)

Content	Included n (%)	Not included n (%)	Planned n (%)
Formulating clinical questions (PICO)	20 (95.2)	1 (4.8)	–
Literature searching	20 (95.2)	1 (4.8)	–
Critical appraisal of research	16 (76.2)	5 (23.8)	–
Appraisal/use of clinical practice guidelines	13 (61.9)	8 (38.1)	–
Levels of evidence/strength of recommendation	17 (81.0)	4 (19.0)	–
AI-based evidence summarization	4 (19.0)	10 (47.6)	7 (33.3)

AI-assisted evidence summarization and search efficiency, while 47.6% did not include it and 33.3% were planning or improving its adoption.

Teaching Methods and Assessment

Lecture was the predominant method (90.5%), followed by clinical-practice-based EBM activities (52.4%), team-based learning (33.3%), and problem-based learning (28.6%), with journal clubs used by only 9.5% (Table 4). AI/digital tools were not used in the EBM process at 57.1% of institutions. Assessment methods (multiple response) were assignments (CAT/appraisal, 81.0%), presentations (71.4%), and written examinations (47.6%); however, EBM assessment results were reflected in graduation requirements or grades at only 38.1%, and 33.3% did not reflect them.

Theory–Clinical Integration and Personnel

Bedside EBM learning based on actual patient questions during clerkship was most often left to preceptor discretion (57.1%), with regular or partial implementation in 19.0% each (Table 4). Linkage between EBM theory faculty and clinical preceptors was based on individual instructors' efforts in

47.6% and was reported as not achieved in 33.3%, with only 19.0% reporting systematic integration within the curriculum. Statisticians or information specialists participated partially in 81.0% of institutions but always in only 9.5%, and 47.6% reported that dedicated EBM personnel were insufficient or very insufficient.

Barriers and Support Needs

The greatest operational difficulty was a shortage of instructors (76.2%), followed by lack of educational materials (28.6%), lack of instructional time (23.8%), and lack of assessment methods (23.8%) (Table 5). The most frequently cited barriers were EBM being deprioritized relative to the heavy volume of medical knowledge (52.4%) and low awareness of the necessity of EBM within the curriculum (52.4%), followed by insufficient clinical faculty EBM teaching capacity (lack of role modeling, 42.9%). Materials considered helpful if provided by a society included a standardized EBM curriculum guide (90.5%), faculty development programs (train-the-trainer [TTT], 76.2%), shared online modules (66.7%), CAT templates/cases (57.1%), and assessment rubrics (52.4%); willingness to use externally developed materials was 100%.

Table 4. Teaching Methods, Assessment, and Theory–Clinical Integration (n=21)

Item	Category	n (%)
Teaching methods (Q15, multiple)	Lecture	19 (90.5)
	Clinical-practice-based EBM activities	11 (52.4)
	Team-based learning (TBL)	7 (33.3)
	Problem-based learning (PBL)	6 (28.6)
	Journal club	2 (9.5)
AI/digital tools (Q16)	Not used	12 (57.1)
	Used (LLM, SR automation, etc.)	9 (42.9)
EBM during clerkship (Q17)	Left to preceptor discretion	12 (57.1)
	Conducted regularly	4 (19.0)
	Conducted partially	4 (19.0)
	Rarely	1 (4.8)
Theory–clinical integration (Q18)	Based on individual instructors' efforts	10 (47.6)
	Not integrated	7 (33.3)
	Systematically integrated in curriculum	4 (19.0)
Assessment methods (Q20, multiple)	Assignment (CAT, appraisal)	17 (81.0)
	Presentation	15 (71.4)
	Written examination	10 (47.6)
Reflected in graduation/grades (Q21)	Yes	8 (38.1)
	Partially	6 (28.6)
	No	7 (33.3)

Table 5. Personnel, Barriers, and Support Needs (n=21)

Item	Category	n (%)
Adequacy of dedicated personnel (Q22)	Moderate	10 (47.6)
	Insufficient	9 (42.9)
	Very insufficient	1 (4.8)
	Sufficient	1 (4.8)
Operational difficulties (Q23, multiple)	Shortage of instructors	16 (76.2)
	Lack of educational materials	6 (28.6)
	Lack of instructional time	5 (23.8)
	Lack of assessment methods	5 (23.8)
	Low student interest	2 (9.5)
Barriers (Q24, multiple)	EBM deprioritized amid heavy content load	11 (52.4)
	Low awareness of EBM's necessity	11 (52.4)
	Insufficient clinical faculty capacity (lack of modeling)	9 (42.9)
	Inadequate IT environment for EBM practice	4 (19.0)
Desired society support (Q25, multiple)	Standardized EBM curriculum guide	19 (90.5)
	Faculty development program (TTT)	16 (76.2)
	Shared online modules	14 (66.7)
	CAT templates/cases	12 (57.1)
	Assessment rubrics	11 (52.4)
Willingness to use external resources (Q26, multiple)	Willing to use materials	21 (100)
	Willing to use programs	9 (42.9)

Governance-Centered Exploration of Structural Factors (Exploratory)

Because the small sample precluded confirmatory testing and causal inference, the relationships between structural factors and outcomes were analyzed descriptively and exploratorily. The following results reflect co-occurrence and should be interpreted as hypothesis-generating.

Operationalizing governance. EBM governance was operationalized as “having a dedicated unit” (Table 1) and “an integrated methodology-clinical design” (Table 2); institutions with at least one of these were classified as having strong governance (15 schools), and those with neither as weak governance (6 schools).

Outcomes differed by the presence of governance. Key outcome indicators differed by strong versus weak governance (Table 6). Systematic theory-clinical integration (27% vs. 0%), inclusion of clinical practice guidelines (80% vs. 17%), reflection in graduation/grades (47% vs. 17%), and EBM conducted during clerkship (47% vs. 17%) were all higher in the strong-governance group, as was mean content coverage (4.60 vs. 2.83). Inclusion of the most basic step (PICO) was high in both groups (100% vs. 83%, near a ceiling effect), whereas higher-order steps such as critical appraisal (87% vs. 50%) and clinical practice guidelines were higher in the strong-governance group. Adding clinical faculty involvement as an additional axis (reverse-coded from the perceived lack of modeling) did not clearly differentiate outcomes; however, because this proxy is an indirect, subjective mea-

sure, whether its weak discrimination reflects a true null or a measurement limitation cannot be determined.

Governance was not explained by exogenous structural characteristics. We explored whether the distribution of governance was associated with schools’ exogenous structural characteristics (Table 7). Admission quota (small 86%, medium 67%, large 62%), establishment type (national 57%, private 79%), region (metropolitan 78%, non-metropolitan 67%), or number of tertiary hospitals (0: 50%, 1: 67%, ≥ 2 : 86%) did not show a consistent association with the proportion having strong governance. If anything, class size was weakly inverse, with a lower mean admission quota in the strong-governance group (78.8) than in the weak-governance group (92.7). In a quota-weighted sensitivity analysis of responding institutions, the distribution of structural factors was likewise independent of school size (weighted vs. unweighted differences mostly within ± 2 percentage points). The number of tertiary hospitals was weakly associated with some clinical-application indicators such as clinical practice guidelines but was unrelated to systematic integration.

In summary, variation in EBM education appeared more related to governance (organization)—a dedicated unit and an integrated design—than to educational content or a school’s outward size, clinical infrastructure, establishment type, or region. However, these are exploratory, correlational observations from a small cross-sectional dataset, and the question of what shapes governance cannot be resolved with these data.

Table 6. Key Outcome Indicators by EBM Education Governance (n/N, %) — Exploratory Descriptive Analysis

Outcome indicator	Strong governance (n=15)	Weak governance (n=6)
Systematic theory-clinical integration	4/15 (27)	0/6 (0)
Inclusion of clinical practice guidelines	12/15 (80)	1/6 (17)
Reflected in graduation/grades	7/15 (47)	1/6 (17)
EBM conducted during clerkship	7/15 (47)	1/6 (17)
Inclusion of critical appraisal	13/15 (87)	3/6 (50)
Inclusion of clinical questions (PICO)	15/15 (100)	5/6 (83)
Content coverage (0–5, mean)	4.60	2.83
No. of teaching methods (mean)	2.40	1.83

Strong governance = having a dedicated unit or an integrated methodology-clinical design (15 schools); weak = neither (6 schools). Values are the number of institutions meeting the indicator / total within the subgroup (%). Because of the small sample, confidence intervals are wide; results are exploratory and cannot be interpreted causally.

Table 7. Proportion with Strong Governance by Exogenous Structural Characteristics (N/N, %) — Exploratory Descriptive Analysis

Exogenous characteristic	Category	Strong governance n/N (%)
Admission quota (2026)	Small (≤ 49)	6/7 (86)
	Medium (50–99)	4/6 (67)
	Large (≥ 100)	5/8 (62)
Establishment type	National	4/7 (57)
	Private	11/14 (79)
Region	Seoul Metropolitan Area	7/9 (78)
	Non-metropolitan	8/12 (67)
No. of tertiary hospitals	None (0)	1/2 (50)
	1	8/12 (67)
	≥ 2	6/7 (86)

Mean admission quota was 78.8 in the strong-governance group vs. 92.7 in the weak-governance group. Admission quota is based on the 2026 academic year (reverted to 3,058); tertiary (advanced general) hospitals are based on the MOHW 5th-cycle (2024–2026) designation. No exogenous characteristic showed a consistent association with the proportion having strong governance (exploratory, not causal).

Discussion

This cross-sectional survey of 21 Korean medical schools found that, although EBM education is universally mandatory, substantial unevenness exists in operational infrastructure, theory–clinical integration, faculty capacity, assessment, and AI integration.

First, EBM education showed a “universal but structurally fragile” pattern. All institutions offered EBM as mandatory and coverage of basic steps such as PICO and literature searching was high, yet 42.9% operated without a dedicated unit and delivery was often integrated and dispersed. This suggests that although EBM has become established as content, the governance responsible for coordinating it remains immature. Given that ASK2026 requires a designated integrated-education lead faculty member and an education committee as the operating body of the integrated curriculum [5], clarifying the organizational responsibility for EBM is a priority.

Second, vertical integration linking theory and clinical practice emerged as the weakest point. That bedside EBM depended on preceptor discretion (57.1%) and that only 19.0% reported systematic theory–clinical integration indicate a disconnect in which EBM remains as early-year methodology teaching and fails to extend to the clinical setting. This reveals a gap between practice and both ASK2026’s requirement for basic–clinical vertical integration (K.2.3.2) and proposals to place EBM as a longitudinal spiral track in the integrated 6-year curriculum [6,7].

Third, the absence of faculty capacity and role modeling was a key barrier. A shortage of instructors was the greatest operational difficulty (76.2%), insufficient clinical faculty EBM teaching capacity (42.9%) was cited as a major barrier, and demand for faculty development programs (TTT) was high (76.2%). Given the importance of faculty who practice EBM at the bedside as role models for students’ EBM attitudes, developing EBM teaching skills for clinical faculty is needed.

Fourth, integration of AI/digital tools remained at an early stage. More than half (57.1%) did not use AI tools and AI-based evidence summarization was included in only 19.0%, although 33.3% reported planning or improving its adoption. As the use of large language models in evidence retrieval and summarization spreads, education should address both the efficiency of use and a critical stance toward the verification, accountability, and limitations of evidence.

Fifth, demand for society-level standardization support was virtually unanimous. Demand for a standardized curriculum guide (90.5%) and willingness to use external materials

(100%) were especially high, indicating that, rather than relying on individual schools’ capacity, provision by relevant societies of standardized curricula, assessment rubrics, shared online modules, and faculty development programs would meet field needs.

Sixth, when structural factors were reanalyzed with a focus on governance, key outcomes (systematic integration, clinical practice guidelines, reflection in assessment, and clerkship EBM) differed by the presence of governance operationalized as a dedicated unit and an integrated design. In contrast, exogenous structural characteristics such as class size, number of tertiary hospitals, establishment type, and region showed no clear association with the distribution of governance or with outcomes, and class size was even weakly inverse [10]. This suggests that variation in EBM education may be more related to organizing EBM under clear responsibility and an integrated design than to educational content or outward resource size; that is, larger or more clinically resourced institutions did not necessarily have better-organized EBM education. However, these are exploratory, correlational observations from a small cross-sectional dataset and cannot serve as a basis for causal inference or for identifying the determinants of governance.

Limitations

This study has several limitations. First, 21 of the 40 Korean medical schools responded, covering 56.8% of the national admission quota; however, because national and Seoul metropolitan area schools responded at relatively higher rates, non-response bias cannot be excluded. Second, the small sample and subgroup sizes (e.g., only 6 schools in the weak-governance group) yield wide confidence intervals and limited statistical power; therefore, the governance and exogenous-characteristic analyses were exploratory and hypothesis-generating, without correction for multiple comparisons or multivariable modeling. Third, governance and clinical faculty involvement were indirectly operationalized from questionnaire items and thus have measurement limitations, and the determinants of governance could not be identified. Fourth, the study relied on a single self-reported respondent per institution, and its cross-sectional design precludes causal or temporal inference. Future multi-institution, multi-respondent studies combining curricular document analysis and student outcome measures are needed.

EBM education in Korean medical schools is largely established in terms of content, being universally mandatory with high coverage of the classic core steps. However, substantial variation exists across institutions in dedicated organization, theory–clinical vertical integration, faculty capacity,

meaningful reflection in assessment, and AI integration. In exploratory analysis, this variation was associated with the presence of governance (a dedicated unit and an integrated design) and was not explained by exogenous structural characteristics such as class size, clinical infrastructure, establishment type, or region. This suggests that the quality of EBM education may be more related to organization than to resource size. During the transition to the integrated 6-year curriculum and ASK2026, it will be important to redesign EBM as a longitudinal competency running through the 6-year curriculum and to provide, at the society level, support for organization (dedicated responsibility structures and integrated design) alongside standardized curricula, assessment tools, and faculty development. However, as a small cross-sectional survey, this study cannot establish causality.

Conflict of Interest

Hyun Jung Kim has been a editor of Journal of Evidence-Based Practice since 2025, and Soo Young Kim has been a editor of Journal of Evidence-Based Practice since 2025. However, they were not involved in the peer reviewer selection, evaluation, or decision process of this article. No other potential conflicts of interest relevant to this article were reported.

Funding

None.

Data Availability Statement

All the data is contained in the manuscript.

Ethics Approval and Consent to Participate

Not applicable.

Authors' Contributions

Conceptualization: HJK JHJ SYK. Data curation: HJK SYK. Formal analysis: HJK SYK MO. Methodology: HJK JHJ MO GJC SYK. Project administration: HJK HS. Visualization: HJK HS MO. Writing – original draft: HJK HS SYK JHJ. Writing – review & editing: HJK JHJ MO GJC SYK HS.

Acknowledgments

None.

ORCID

Hyun Jung Kim, <https://orcid.org/0000-0003-2018-2385>
Mookyung Oh, <https://orcid.org/0000-0002-2011-5708>
Jae Hung Jung, <https://orcid.org/0000-0002-4990-7098>
Geun Joo Choi, <https://orcid.org/0000-0002-4653-4193>
Heather Swan, <https://orcid.org/0000-0001-9129-5008>
Soo Young Kim, <https://orcid.org/0000-0002-3205-9408>

References

1. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. *BMJ* 1996; 312: 71-2.
2. Evidence-Based Medicine Working Group. Evidence-based medicine: a new approach to teaching the practice of medicine. *JAMA* 1992; 268: 2420-5.
3. Dawes M, Summerskill W, Glasziou P, Cartabellotta A, Martin J, Hopayian K, et al. Sicily statement on evidence-based practice. *BMC Med Educ* 2005; 5: 1-7.
4. Ministry of Education. Enforcement Decree of the Higher Education Act [Internet]. Sejong: Ministry of Education; 2023 [cited 2026 08 15]. Available from <https://www.law.go.kr/>
5. Korean Institute of Medical Education and Evaluation. [Accreditation standards for medical education]. Seoul (Korea): Korean Institute of Medical Education and Evaluation; 2025. Available from: <https://kimee.or.kr/board/data/?pageid=1&mod=document&category1=%EA%B8%B0%EC%A4%80&uid=1739>. Accessed September 23, 2026.
6. Kim Y. The integrated 6-year medical curriculum in Korea: the beginning of a journey toward change. *Korean Med Educ Rev* 2025; 27: 97-9.
7. Lee SH, Ahn J, Lee MJ, Ha CW. Development and operation of a learner-centered integrated 6-year medical program. *Korean Med Educ Rev* 2025; 27: 132-43.
8. Kim K. A case of student-centered integrated 6-year medical curriculum development: early implementation and challenges. *Korean Med Educ Rev* 2025; 27: 154-61.
9. Yoo DM, Kang WS. A practical approach: reform of the 6-year medical curriculum at the Catholic University of Korea. *Korean Med Educ Rev* 2025; 27: 162-8.
10. Ministry of Health and Welfare. Designation of 5th-cycle (2024-2026) tertiary (advanced general) hospitals [Internet]. Sejong: Ministry of Health and Welfare; 2025 [cited 2026 08 15]. Available from <https://www.mohw.go.kr/>

Variability, algorithm conformance, and accuracy of large language model–based tools for risk-of-bias (RoB 2.0) assessment of randomized trials: a pilot study

Jiae Choi^{1,2}, Heather Swan^{1,2}, Min Jung Kim^{1,2}, Hyun Jung Kim^{1,2}

¹Cochrane Korea, Seoul, Korea

²Institute for Evidence-Based Medicine, College of Medicine, Korea University, Seoul, Korea

Background: Large language model (LLM)–based tools are increasingly used to automate risk-of-bias assessment of randomized controlled trials with the revised Cochrane tool (RoB 2.0). However, their reproducibility, their accuracy, and their fidelity to the deterministic algorithm mapping signalling-question responses to domain judgments remain unclear.

Methods: In a conference workshop, participants used an identical prompt and tool versions to assess one RCT with three configurations and entered each tool's output verbatim (13, 15, and 10 runs). One experienced reviewer's assessment served as the reference. For each run we computed run-to-run reproducibility, agreement with the expert, and conformance between the tool's stated domain judgment and the judgment implied by applying the RoB 2.0 algorithm to that run's own signalling answers.

Results: All configurations showed substantial run-to-run variability under identical conditions, greatest in the conditionally complex domain 2 (Gemini pairwise agreement 0.34). Expert agreement varied widely across configurations (mean 5–60%), and one configuration systematically under-rated risk. Even in domains with a fully specified algorithm, stated judgments frequently diverged from the value implied by the tool's own signalling answers (domain 2, mean 42%). Skip-logic violations occurred in 69%, 80%, and 100% of runs. Recomputing domain judgments from the signalling answers improved expert agreement for the general-purpose configurations.

Conclusion: LLM-based RoB 2.0 assessments exhibited variability and errors. Whatever tool is adopted, its characteristics and variability must be recognized. Having the tool perform only atomic (single) judgments while delegating aggregation to the algorithm, together with human review, may improve accuracy.

Keywords: Large language models; Reproducibility of results; Artificial intelligence; Systematic reviews as topic; Randomized controlled trials as topic

Introduction

Risk-of-bias assessment of randomized controlled trials (RCTs) is a core step in systematic reviews and in judging the certainty of evidence (GRADE). The revised Cochrane tool, RoB 2.0, has a hierarchical structure: reviewers answer structured signalling questions across five domains, map those answers to a domain-level judgment (low, some concerns, high) via a formal algorithm, and then combine the domain

judgments into an overall judgment [1,2]. This process demands substantial time and judgment even from trained assessors and is a bottleneck in large reviews [3,4].

To reduce this burden, automation with large language models (LLMs) has been explored. Following early systems such as RobotReviewer, purpose-built RoB 2.0 tools that combine document retrieval, LLM prompting, and human review—such as ROBot2—have recently been proposed [5,6]. However, LLMs generate probabilistic output and may

Received: August 1, 2026 **Accepted:** August 31, 2026

Corresponding author: Hyun Jung Kim

Email: iebm.ku@gmail.com

© 2026 Korean Society of Evidence-Based Medicine

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

respond differently to identical input across runs, and systematic biases arising from alignment training have been reported. How these characteristics affect the reproducibility and accuracy of a task such as RoB 2.0 assessment, which requires multi-step conditional logic and deterministic aggregation, has not been well studied.

Using a design in which a single RCT article was repeatedly assessed with an identical prompt and identical tool versions, we quantified (1) run-to-run reproducibility, (2) accuracy relative to an expert reference, and (3) conformance to the algorithm mapping signalling answers to domain judgments. We further examined whether recomputing domain judgments from a tool's own signalling answers improves accuracy. We studied three configurations that provide the manual in different ways (full manual in-context, manual via retrieval, and a purpose-built pipeline); rather than ranking these approaches, we focused on characterizing the variability common to all and the error patterns specific to each.

Methods

Study Design and Data

At a workshop session of the 3rd conference of the Korean Society for Evidence-Based Medicine, participants used an identical prompt and identical versions of three LLM-based configurations—general-purpose Gemini, a custom Gemini Gem, and ROBot2 (Roboto)—to assess one RCT article immediately, entering each tool's output (signalling-question responses and domain judgments) verbatim. Thirteen, fifteen, and ten independent assessments were collected, respectively (Table 1). Because the prompt, tool, and target article were fixed, differences across repeated assessments reflect each configuration's run-to-run variability.

The three configurations differed in how the RoB 2.0 manual (the detailed tool guidance from the official website) was delivered, and in architecture. Gemini (general) and Gemini Gems used the same base model, the same manual, and the same prompt, differing only in manual delivery: Gemini (general) had the full manual pasted into the chat context so that the entire manual was continuously available

to the model, whereas Gemini Gems attached the manual as reference knowledge in a custom Gem. A Gem's reference knowledge is handled by retrieval-augmented generation (RAG), surfacing only passages relevant to each query. Roboto (ROBot2) is a purpose-built RoB 2.0 pipeline that combines document retrieval and LLM prompting with the official flowchart aggregation logic. The three configurations thus correspond to full-manual-in-context, manual-via-RAG, and a purpose-built pipeline; our aim was not to rank these approaches but to characterize the variability and error patterns arising in each.

One experienced reviewer's assessment served as the reference. During initial cross-checking, three transcription errors in the expert's entries (signalling questions 3.3, 5.1, 5.2) were detected via an algorithm-conformance check and corrected; after correction, the expert's assessment matched the algorithm-implied judgment in all five domains (Table 1).

RoB 2.0 Algorithm Implementation

We implemented in code the RoB 2.0 algorithm mapping signalling answers to domain judgments. Domains 1, 2, and 5 were fixed using the official guidance decision tables [2]. For domains 3 and 4, the signalling-question structure differs between the 2016 draft and the 2019 final version (the 2016 domain 3 has three questions and domain 4 two, whereas the 2019 final version has four and five, respectively); because our data and the studied tools use the 2019 final structure, we encoded accordingly, and the expert reference matching all five domains provided empirical validation. Typographic variants of the domain-judgment labels were standardized at the analysis stage (raw data unchanged).

Analysis Metrics

Reproducibility. Run-to-run reproducibility per domain was quantified with the Simpson pairwise-agreement index ($\Sigma\pi_2$; the probability that two randomly drawn assessments give the same judgment) and normalized Shannon entropy.

Accuracy. The agreement between each tool's domain judgment and the expert reference was computed per domain.

Table 1. Study Sample and Configurations

Configuration	Runs (n)	Domain-judgment source	Manual delivery
Gemini (general)	13	LLM-generated	Full manual in-context
Gemini Gems	15	LLM-generated	Manual as RAG knowledge (Gem)
Roboto (ROBot2)	10	Algorithm (flowchart), as designed	Purpose-built RoB 2.0 pipeline
Expert reference	1	Human judgment	Algorithm-conformant (5/5, verified)

LLM, large language model; RAG, retrieval-augmented generation. Roboto is designed to produce judgments by applying the official flowchart to the signalling answers.

Algorithm conformance and skip-logic. For each run, we computed the agreement between the judgment implied by applying the algorithm to that run's own signalling answers and the judgment the tool actually entered. We also counted violations of the skip-logic by which subsequent questions should be marked "not applicable" depending on earlier answers.

Recomputation. Setting aside the tool's entered judgment, we recomputed the judgment by applying the algorithm to each tool's own signalling answers and evaluated the change (Δ) in expert agreement.

Statistical Analysis

All analyses were descriptive, consistent with the pilot nature of the study. Domain judgments were coded ordinally (low = 0, some concerns = 1, high = 2). Run-to-run reproducibility was quantified with the Simpson pairwise-agreement index ($\Sigma\pi^2$; closer to 1 indicates greater consistency) and normalized Shannon entropy (0 = complete agreement, 1 = maximum dispersion). Because the design is repeated measurement of a single result (one article), Cohen and Fleiss kappa (κ), which presuppose multiple assessment targets, are by definition inappropriate and were not used.

Accuracy relative to the expert and algorithm conformance were computed as domain-level percent agreement. Skip-logic violations were summarized as the number of violations per run and the proportion of runs with ≥ 1 violation. The recomputation effect was expressed as the change (Δ) obtained by subtracting the expert agreement of the tool's entered judgment from the expert agreement of the judgment recomputed from the signalling answers.

To convey the precision of the per-configuration summary estimates (five-domain means of reproducibility, expert agreement, and conformance), 95% confidence intervals (CIs) were obtained by nonparametric bootstrap (2,000 resamples, percentile method) resampling runs within each configuration. Formal hypothesis testing of between-configuration differences was not performed, to avoid over-interpretation given the single-target, small-sample pilot design. All analyses were performed in Python (pandas, numpy).

Results

Run-to-Run Reproducibility

Despite identical conditions, all three configurations showed substantial run-to-run variability. The five-domain mean pairwise agreement was 0.76 (95% CI 0.72–0.86) for Roboto, 0.66 (0.61–0.77) for Gems, and 0.59 (0.57–0.69) for Gemini (general) (Table 2; Fig. 1a). Variability differed markedly by domain, being greatest in the conditionally complex domain 2, where Gemini's pairwise agreement of 0.34 approached chance and Roboto's was 0.42. Domains 1 and 5, with simpler signalling-question structures, were comparatively stable.

Accuracy Relative to the Expert Reference

Agreement with the expert reference (domain 1 = some concerns; domains 2–5 = high) varied widely across configurations (five-domain means: Roboto 60%, Gemini 24%, Gems 5%; bootstrap 95% CIs 50–70%, 14–35%, 1–11%, respectively; Table 2). In particular, Gems combined moderate reproducibility with only 5% expert agreement, reflecting a tendency to rate most domains as low risk and thus diverging system-

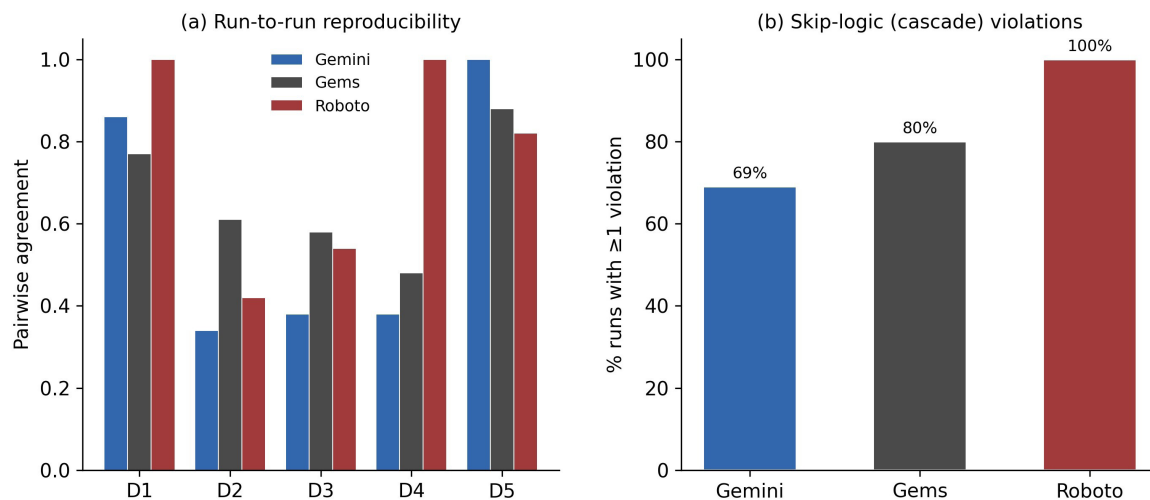


Fig. 1. (a) Run-to-run reproducibility (pairwise agreement) by domain. (b) Proportion of runs with ≥ 1 skip-logic (cascade) violation, by configuration.

atically from the expert's high-risk judgments. Reproducibility (precision) and accuracy (closeness to the reference) were thus independent, and Gems showed a "reproducibly under-rating" directional bias (Fig. 2).

Algorithm Conformance and Internal Consistency

Although the mapping from signalling answers to judgments is deterministic, stated judgments frequently diverged from the implied values even in domains with fully specified official tables. Conformance was 68% for domain 1 and 42% for domain 2 (mean across configurations); Gems was as low as 27% in domain 2 and 13% in domain 4, whereas the expert reference matched in 100% of both domains (Table 3). The five-domain mean conformance was 43% (95% CI 29–58%) for Gemini, 51% (39–63%) for Gems, and 60% (40–78%) for Roboto. This is an error source distinct from stochastic variability, indicating that tools generate plausible signalling answers yet fail to faithfully execute the fixed aggregation rule.

Skip-logic conformance also revealed structural problems. The proportion of runs with ≥ 1 violation was 69% (mean 2.5 per run) for Gemini, 80% (2.9) for Gems, and 100% (4.3) for Roboto; the expert reference had none (Fig. 1b). Roboto in particular violated the skip-logic in every run, producing answers to all subsequent questions regardless of the preceding results. Moreover, in Roboto some entered judgments did not follow from the tool's own answers; for example, a run

Table 2. Domain- Level Reproducibility and Agreement with The Expert Reference, by Configuration

Domain	Gemini		Gems		Roboto	
	Repro.	Agree.	Repro.	Agree.	Repro.	Agree.
D1	0.86	8%	0.77	13%	1.00	0%
D2	0.34	38%	0.61	0%	0.42	40%
D3	0.38	38%	0.58	7%	0.54	70%
D4	0.38	38%	0.48	7%	1.00	100%
D5	1.00	0%	0.88	0%	0.82	90%
Mean	0.59	24%	0.66	5%	0.76	60%

Repro., run-to-run reproducibility (Simpson pairwise agreement, 0–1; higher = more consistent). Agree, percent agreement of the domain judgment with the expert reference. D1–D5 = domains 1–5.

Table 3. Algorithm Conformance, Skip-Logic Violations, and Recomputation Effect, by Configuration

Configuration	Conformance (mean)	Runs with ≥ 1 violation	Violations/run	Recompute Δ (mean)
Gemini (general)	43%	69%	2.5	+0.34
Gemini Gems	51%	80%	2.9	+0.24
Roboto	60%	100%	4.3	Uninterpretable*
Expert reference	100%	0%	0	—

Conformance, proportion of runs in which the entered judgment matched the value implied by the run's own signalling answers (five-domain mean). *For Roboto the recomputation effect is not interpretable because of answer-judgment decoupling (see Results 3.4).

answering "no information" to all three signalling questions of domain 1 had an entered judgment of "low risk," which no correct algorithm can yield when all answers are "no information" (the official rule gives "some concerns"). Because participants transcribed the tool's output verbatim, this indicates that the judgment displayed by the tool was not coherent with the answers it displayed—i.e., the answer-judgment link was broken at the tool level.

Effect of Recomputing Judgments from Signalling Answers

Setting aside the tool's entered judgment, we recomputed the domain judgment by applying the algorithm to each tool's own signalling answers and assessed the change in expert agreement. For the two general-purpose configurations, whose judgments are generated directly by the LLM, recomputation clearly improved accuracy. Gemini improved in all five domains (mean Δ +0.34; domain 4 +0.54, domain 2 +0.39), and Gems improved substantially in domains 1, 2, and

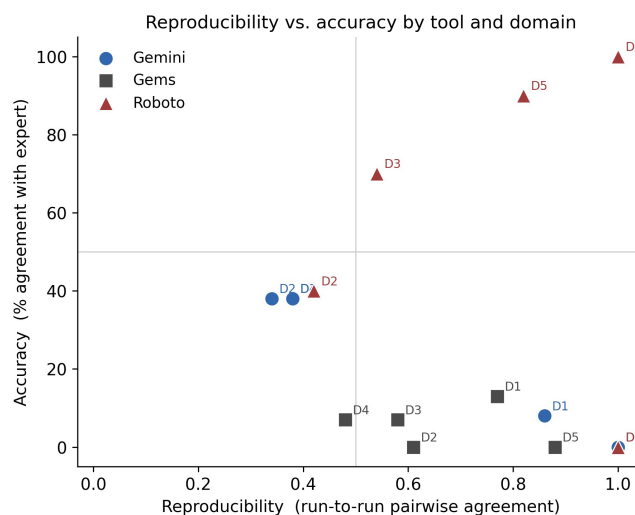


Fig. 2. Reproducibility (x-axis) versus accuracy relative to the expert reference (y-axis) for each configuration and domain. The upper-right region is ideal; the lower-right region ("reproducibly inaccurate") is the most concerning.

4 (domain 4 +0.66, domain 2 +0.40; Fig. 3). This suggests that these configurations' errors arise mainly at the aggregation step rather than in the signalling answers themselves, with the largest gains in the conditionally complex domains 2 and 4.

For Roboto, recomputation tended to lower expert agreement (domains 3, 4, 5), but this result requires caution. Because Roboto's domain judgment is by design produced by applying the official flowchart to the signalling answers, the very fact that recomputation changes the judgment reflects the answer-judgment decoupling described above. The premise of the recomputation comparison therefore does not hold for this tool, and its recomputation effect is not an interpretable quantity; accordingly, the finding of improved accuracy through recomputation holds only for the configurations that generate judgments directly.

Discussion

This pilot study, using a repeated-assessment design under identical conditions, identified three limitations of current

LLM-based RoB 2.0 assessment.

First, structural limitation. The tools did not faithfully execute RoB 2.0's conditional logic and deterministic aggregation. Skip-logic violations were pervasive (Roboto violated them in every run), and even in domains with a fully specified official algorithm, entered judgments frequently diverged from the value implied by the tool's own answers. This reveals a deficit in executing multi-step rules as specified, separate from the ability to generate answers. Notably, Gemini (general) and Gemini Gems were given the same official manual yet showed low conformance and accuracy, indicating that the limitation is not a matter of access to the criteria but of faithfully executing the deterministic multi-step logic.

Second, stochastic variability. Judgments scattered across runs despite identical input, especially in conditionally complex domains (near chance in domain 2). This is a serious constraint for an assessment tool that requires reproducibility.

Third, systematic directional bias. The general-purpose configurations here under-rated risk (Gems expert agree-

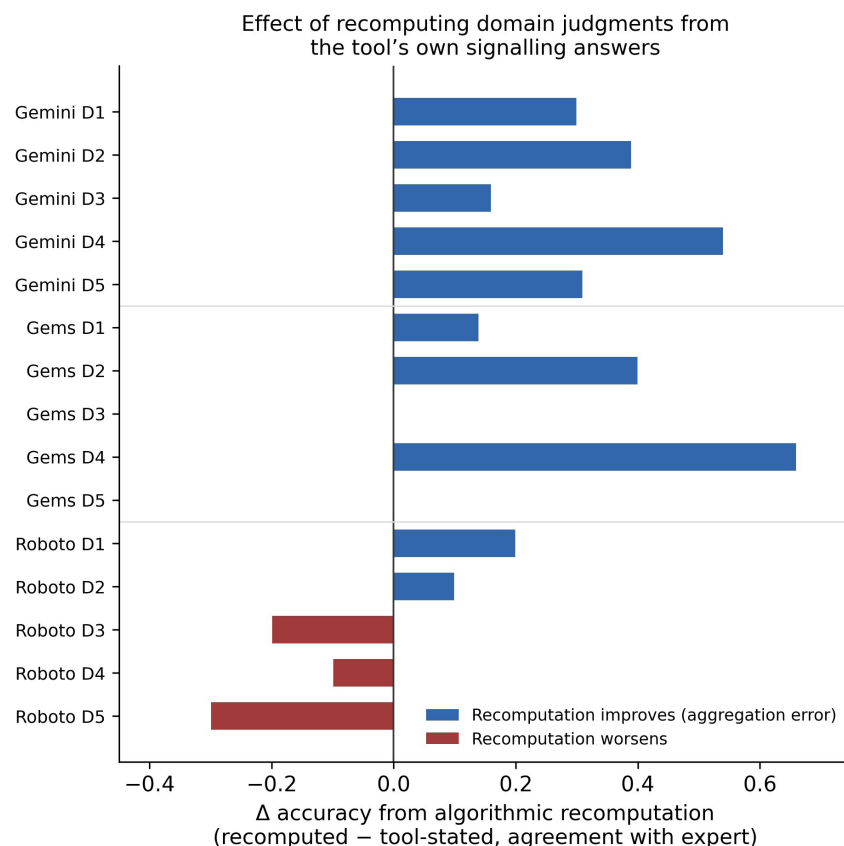


Fig. 3. Change in expert agreement (Δ) when domain judgments are recomputed from each tool's own signalling answers. Positive values indicate improvement (aggregation error); negative values indicate worsening. For Roboto the negative values reflect answer-judgment decoupling and are of limited interpretability.

ment 5%). The direction of bias, however, depends on the tool and its tuning; a purpose-built pipeline has conversely been reported to be overly conservative [5]. It is therefore more appropriate to say that a systematic directional bias exists and that its direction differs across tools, rather than to generalize a single “optimistic bias.”

Although differing in approach, all three configurations showed substantial run-to-run variability and errors, with somewhat different problem profiles. The configuration providing the full manual in-context (Gemini) had large variability in complex domains and frequently failed at the step of aggregating answers into judgments. The configuration providing the same manual as RAG reference knowledge (Gems) showed a pronounced tendency to under-rate risk, consistent with weaker grounding when only selected passages are surfaced. The purpose-built pipeline (Roboto) had relatively higher reproducibility and accuracy in some domains but showed structural incoherence—violating the skip-logic in every run and displaying judgments that did not follow from its answers. In short, no approach was free of variability and error, and each had its own characteristic weakness.

The implication of this study therefore does not lie in ranking a particular delivery mechanism or architecture. The key point is that, whatever tool is used in the future, its characteristic behavior, variability, and typical error patterns must be adequately recognized and anticipated. When adopting LLM-based RoB assessment, rather than trusting a tool without verification, checking variability through repeated assessment, verifying the consistency between signalling answers and judgments (algorithm conformance), and combining human review with algorithmic aggregation should be standard safeguards.

These findings have practical implications. As the recomputation analysis showed, much of the error in general-purpose configurations arises not in the signalling answers themselves but in the composite step of aggregating them into judgments. Rather than delegating composite, multi-step judgments to the tool, a hybrid approach in which the tool performs only atomic (single) signalling judgments while the deterministic aggregation is handled algorithmically (in code) may improve both accuracy and reproducibility. Indeed, recomputation from signalling answers alone substantially improved domain accuracy for the general-purpose configurations.

Some limitations (the accuracy of the signalling answers) are expected to improve as model capability advances. However, stochastic variability and the faithful execution of deterministic multi-step logic are more robustly addressed by architectural design—separating human-in-the-loop re-

view from algorithmic aggregation—than by awaiting larger models. In addition, our algorithm-conformance check functioned as a data-quality audit, detecting three transcription errors in the expert reference.

Limitations

This is a pilot, hypothesis-generating study based on a single article, a single expert reference, and three configurations (38 assessments in total); generalization requires caution. Given the single-target repeated-measures design, kappa indices were not applied, and the results are specific to particular tool versions and time points. The precise cause of Roboto’s answer–judgment decoupling (implementation, defaults, or display) requires further examination of the tool itself. Larger studies with multiple articles, multiple experts, and various tools and versions are warranted.

Conclusion

Regardless of delivery mechanism or architecture, current LLM-based RoB 2.0 assessment showed stochastic variability, structural non-conformance to deterministic logic, and tool-dependent directional bias, with a distinct error profile for each approach. Although these limitations will be gradually mitigated as AI capability improves, at present the characteristics and variability of whatever tool is used must be heeded. Rather than delegating composite judgments to the tool, it is preferable to focus on the accuracy of single judgments and to combine algorithmic aggregation with human review.

Conflict of Interest

Hyun Jung Kim has been a editor of Journal of Evidence-Based Practice since 2025. However, she was not involved in the peer reviewer selection, evaluation, or decision process of this article. No other potential conflicts of interest relevant to this article were reported.

Funding

The authors received no specific funding for this work.

Data Availability Statement

Not applicable.

Ethics Approval and Consent to Participate

Not applicable.

Authors' Contributions

Conceptualization: HJK. Data curation: HJK JC. Formal analysis: HJK JC HS. Methodology: HJK HS JC. Project administration: MJK HS. Visualization: JC HS MJK. Writing – original draft: HS MJK. Writing – review & editing: HJK JC HS MJK.

Acknowledgments

None.

ORCID

Jiae Choi, <https://orcid.org/0000-0003-4698-7608>

Heather Swan, <https://orcid.org/0000-0001-9129-5008>

Min Jung Kim, <https://orcid.org/0000-0002-6459-8721>

Hyun Jung Kim, <https://orcid.org/0000-0003-2018-2385>

References

1. Higgins JP, Savovic J, Page MJ, Sterne JA. Revised Cochrane risk-of-bias tool for randomized trials (RoB 2) [Internet]. 2019 Aug 22 [cited 2026 08 15]. Available from <https://www.riskofbias.info/welcome/rob-2-0-tool/current-version-of-rob-2>
2. Sterne JA, Savovic J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ* 2019; 366: l4898.
3. Crocker TF, Lam N, Jordao M, Brundle C, Prescott M, Forster A, et al. Risk-of-bias assessment using Cochrane's revised tool for randomized trials (RoB 2) was useful but challenging and resource-intensive: observations from a systematic review. *J Clin Epidemiol* 2023; 161: 39-45.
4. Minozzi S, Cinquini M, Gianola S, Gonzalez-Lorenzo M, Banzi R. The revised Cochrane risk of bias tool for randomized trials (RoB 2) showed low interrater reliability and challenges in its application. *J Clin Epidemiol* 2020; 126: 37-44.
5. Hevia A, Chintalapati S, Lai VKW, Nguyen TT, Wong WT, Klassen TP, et al. ROBot2: an interactive system and dataset for LLM-assisted clinical trial risk of bias assessment. In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*; Stroudsburg (PA): Association for Computational Linguistics; 2025; pp 12-25.
6. Marshall IJ, Kuiper J, Wallace BC. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. *J Am Med Inform Assoc* 2016; 23: 193-201.

Interpreting patient values and preferences in clinical practice guideline development: distinguishing reliance on physician judgment from treatment outcome preferences – an exploratory multicenter study in patients with anal cancer

Hyo Seon Ryu¹, Jung-Myun Kwak¹, Kil-Yong Lee², Dalyong Kim³, Dong Hyun Kang⁴, Heather Swan^{5,6}, Hyun Jung Kim^{5,6}

¹Division of Colon and Rectal Surgery, Department of Surgery, Korea University College of Medicine, Seoul, Korea

²Department of Surgery, Uijeongbu St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Uijeongbu, Korea

³Division of Hematology and Oncology, Department of Internal Medicine, Dongguk University Medical Center, Goyang, Korea

⁴Department of Surgery, Soonchunhyang University Cheonan Hospital, Soonchunhyang University College of Medicine, Cheonan, Korea

⁵Cochrane Korea, Seoul, Korea

⁶Institute for Evidence-Based Medicine, College of Medicine, Korea University, Seoul, Korea

Background: Patient values and preferences are important in clinical practice guideline development, but responses indicating reliance on physician judgment may be difficult to interpret. We examined whether such reliance can coexist with explicit preferences about treatment outcomes and trade-offs.

Methods: We conducted an exploratory secondary analysis of a 22-item cross-sectional survey completed by 14 adults with anal cancer recruited from four hospitals during guideline development. The survey assessed decision-making participation, responses to specific clinical scenarios, rankings of six treatment goals, and treatment-acceptance thresholds. Ten scenario items included the option "I would follow my treating physician's judgment." Analyses were descriptive and emphasized within-participant response patterns.

Results: Ten participants (71.4%) generally preferred shared decision-making, yet 11 selected physician judgment in at least one scenario. All 11 completed a full ranking of treatment goals, and 9 completed both trade-off threshold items. Among seven participants who relied on physician judgment for salvage treatment, all completed the ranking and six completed both thresholds; five required at least a 20% survival improvement to accept surgery resulting in a permanent stoma. Although permanent stoma ranked fourth among six goals overall, 11 of 13 respondents required at least a 20% survival improvement to accept such surgery.

Conclusions: Reliance on physician judgment did not necessarily indicate an absence of patient preferences. Guideline-related preference studies should distinguish preferences for the decision process, general priorities among outcomes, and context-specific benefit-risk trade-offs. Responses indicating reliance on physician judgment should not automatically be treated as "no preference" or missing information.

Keywords: Patient Preference; Practice Guideline as Topic; Shared Decision-Making; Patient Participation; Anus Neoplasms; Surveys and Questionnaires

Received: August 16, 2026; **Revised:** September 1, 2026; **Accepted:** September 8, 2026

Corresponding author: Hyun Jung Kim

Email: moole@korea.ac.kr

© 2026 Korean Society of Evidence-Based Medicine

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Introduction

Clinical practice guideline recommendations are intended to reflect not only the best available evidence but also the values patients place on the benefits, harms, and other consequences of the available options. The Grading of Recommendations Assessment, Development and Evaluation (GRADE) Evidence to Decision framework therefore treats the values placed on important outcomes, including variability or uncertainty in those values, as a distinct consideration when moving from evidence to recommendations [1]. Guidance on incorporating patient values and preferences into guideline development similarly emphasizes eliciting the relative importance patients assign to health outcomes [2].

This concept is related, but not identical to, the role a patient wishes to play in decision-making. Shared decision-making involves information exchange, deliberation, and elicitation of patient preferences, but patients may differ in how much control they want over the final choice [3,4]. Preferences for decisional control can also vary across clinical contexts. Scenario-based research has shown that a general preference for an active, shared, or physician-led role does not necessarily predict the role preferred in a specific decision [5], and studies in cancer care have found that decision characteristics influence whether decisions are patient controlled, shared, or physician controlled [6]. Patients may also value participation and information while still placing substantial weight on a clinician's recommendation [7].

These distinctions create an interpretive problem for guideline developers. When a patient responding to a values-and-preferences survey selects "I would follow my treating physician's judgment," the response may be interpreted as indicating no preference, uncertainty, or inability to choose between options. However, the same patient may have clear views about which outcomes matter, harms they wish to avoid, or the magnitude of benefit or risk that would make a treatment acceptable. Treating reliance on physician judgment as equivalent to an absence of values may therefore discard information that is directly relevant to recommendation development.

A second measurement issue is that "outcome preference" itself can be elicited in different ways. Asking patients to rank outcomes in general identifies their relative priorities, whereas asking how much benefit or harm they would accept in a specific clinical choice elicits a contextual trade-off. These forms of preference information may be related but need not be interchangeable. This distinction is especially relevant to guideline development, where judgments about the balance of desirable and undesirable effects depend not only on

whether an outcome is important but also on how patients value the magnitude of competing consequences.

Anal cancer provides a useful clinical context in which to examine these issues because management includes decisions with markedly different purposes and trade-offs, including screening in high-risk groups, staging and response assessment, radiation dose and systemic therapy, local excision in selected early-stage disease, and salvage surgery that may result in a permanent stoma [8-10]. During the development of an anal cancer clinical practice guideline, we conducted a patient values-and-preferences survey that included a general question about preferred decision-making participation, multiple scenario-specific items, rankings of treatment goals, and treatment-acceptance thresholds. Using these data, the present exploratory analysis examined whether reliance on physician judgment in specific scenarios could coexist with explicit outcome priorities and trade-off preferences, and whether these different forms of preference information should be interpreted separately in guideline development.

Methods

Study design and participants

This study was an exploratory secondary analysis of cross-sectional survey data on patient values and preferences collected during the development of a clinical practice guideline for anal cancer. Eligible participants were adults aged 19 years or older who had been diagnosed with and treated for anal cancer at one of four participating hospitals: Korea University Anam Hospital, Soonchunhyang University Hospital Cheonan, the Catholic University of Korea Uijeongbu St. Mary's Hospital, and Dongguk University Hospital. The survey was conducted from February 11 to July 27, 2026. Eligible patients who visited the participating hospitals during the survey period were approached in the outpatient setting, and the questionnaire was administered face-to-face by an outpatient-clinic nurse. A total of 14 patients participated. Given the rareness of anal cancer and practical constraints on recruitment, convenience sampling was used. The analysis was designed to describe within-participant response patterns rather than to test a prespecified statistical hypothesis or estimate population proportions with precision.

Survey instrument

The survey consisted of 22 items: four items on demographic and clinical characteristics, two general items on the preferred decision maker and priorities among treatment goals, and 16 items corresponding to key questions in the anal

cancer clinical practice guideline. Major clinical scenarios included screening in high-risk populations; additional FDG-PET/CT for staging; timing and use of magnetic resonance imaging for treatment-response assessment; radiation dose in early anal cancer; factors considered when selecting chemotherapy; local excision versus chemoradiotherapy for stage I anal cancer; and salvage treatment for suspected residual cancer.

General decision-making participation was assessed using four response options: “It is best for the physician to make all decisions,” “It is best for the physician and me to discuss and decide together,” “It is best to discuss with my family and follow my family’s opinion,” and “It is best for me to make all decisions.” Participants were also asked to rank six treatment goals from most to least important: cure, symptom relief, quality of life, duration of time during which usual activities could be maintained, treatment-related adverse effects, and whether a permanent stoma would be required. Lower rank values therefore indicated greater importance.

For each clinical scenario item, expected benefits and harms were presented before the patient selected a response. Ten of the 22 items included the option “I would follow my treating physician’s judgment.” For this secondary analysis, selection of this option was operationally coded as reliance on physician judgment. This code was not intended to imply complete transfer of decision-making authority or a desire to avoid participation. Two items separately offered both the physician-judgment option and the option “Both choices have advantages and disadvantages, so it is difficult to decide.” One additional screening-frequency item combined “I do not know” and “I would follow my physician’s judgment” in a single response option. Because uncertainty and reliance on physician judgment could not be distinguished in that item, it was excluded from the primary physician-reliance analysis and examined only in sensitivity analysis.

The survey also assessed two context-specific trade-off thresholds. One item asked the maximum additional local recurrence risk participants would accept when choosing local excision rather than chemoradiotherapy. A second item asked the minimum survival improvement required to accept surgery for suspected residual cancer when the surgery would result in a permanent stoma and involve operative risk.

Classification of clinical scenario items

The 10 items that included an unambiguous physician-judgment option were classified a priori according to clinical content as five diagnostic/evaluation-related items and five treatment-related items. Diagnostic/evaluation-related items

addressed screening in high-risk populations, additional FDG-PET/CT for staging, timing of treatment-response assessment, factors considered when determining assessment timing, and magnetic resonance imaging for response assessment. Treatment-related items addressed the treatment goal in early anal cancer, choice between low-dose and standard-dose radiotherapy, factors considered when selecting chemotherapy, treatment choice for stage I anal cancer, and salvage treatment for suspected residual cancer. This classification was specified before response results were reviewed.

Analysis

Categorical results are presented as n/N (%), where N is the number of valid responses to each item. Age is presented as the median and range. Given the sample size and exploratory purpose, the analysis was descriptive; confidence intervals, formal hypothesis tests, regression analyses, and inferential group comparisons were not used.

For each participant, we calculated the number of diagnostic/evaluation-related and treatment-related items in which physician judgment was selected and whether at least one such response occurred in each category. Correspondence between general participation preference and scenario-specific responses was described at the participant level. All participants provided a response to each of the 10 physician-reliance items. One participant marked more than one non-physician-judgment option on three of these single-choice items; because physician judgment itself was not selected in those multiply marked responses, they were classified as not selecting physician judgment in the primary analysis. A sensitivity analysis excluded this participant. Missing responses to the two threshold items were reported separately and were not imputed.

Treatment-goal rankings were summarized using mean and median rank and the numbers of participants placing each goal in the first, top two, or bottom two positions. To assess whether physician reliance represented an absence of elicitable preference information, we described completion of the six-goal ranking and both threshold items among participants who did and did not select physician judgment, including the salvage-treatment scenario. We also descriptively contrasted general outcome rankings with corresponding context-specific choices or thresholds. These comparisons were intended to examine whether the two forms of preference information were non-equivalent and complementary, not to establish psychometric construct validity.

Sensitivity analyses examined the effect of including the screening-frequency item that combined “I do not know” with “I would follow my physician’s judgment,” the influence

of the participant with multiple markings, and the distribution of responses across the two threshold items to assess whether clustering in the highest survival-benefit category was compatible with a nonspecific tendency to choose end-point response categories.

Ethical considerations

The original patient values-and-preferences survey was administered anonymously. Participants were informed in advance that their responses would not affect clinical decisions and would be used as reference information for clinical practice guideline development.

Results

Participant characteristics

Of the 14 participants, 12 (85.7%) were women, and the median age was 59.5 years (range, 40–85 years). Eleven participants (78.6%) reported no risk factor placing them in a high-risk group for anal cancer. All participants had received both radiotherapy and chemotherapy. One participant had undergone both local excision and abdominoperineal resection (Table 1).

General participation preference and context-specific reliance on physician judgment

When asked generally who should make decisions about additional tests or treatment options, 10 participants (71.4%) preferred to discuss the decision with their physician and decide together, whereas 4 (28.6%) preferred the physician to make all decisions. No participant preferred to follow the family's opinion or to make all decisions independently.

Despite the predominance of a general preference for joint decision-making, 11 of 14 participants (78.6%) selected the option to follow the physician's judgment on at least one of the 10 clinical scenario items. Among the 10 participants who generally preferred to discuss and decide together, 7 selected the physician-judgment option at least once; their median number of such selections was 1.5 (range, 0–6). All 4 participants who generally preferred the physician to make all decisions selected physician judgment in at least one scenario, with a median of 2.5 selections (range, 1–5).

Across the five diagnostic/evaluation-related items, reliance on physician judgment accounted for 7 of 70 item responses (10.0%), ranging from 0 to 2 participants per item. Across the five treatment-related items, it accounted for 20 of 70 item responses (28.6%), ranging from 3 to 7 participants per item (Table 2). After excluding the salvage-treatment item, 13 of 56 responses (23.2%) across the remaining four treatment-related items indicated reliance on physician judgment. At the participant level, reliance was more frequent for treatment-related than diagnostic/evaluation-related items in 8 participants, equal across the two item types in 4, and more frequent for diagnostic/evaluation-related items in 2. Because the item types also differed in risk, complexity, and clinical consequences, these frequencies were treated as descriptive rather than comparative evidence.

Availability of outcome-preference information despite reliance on physician judgment

All 14 participants completed a full, nonduplicated ranking of the six treatment goals. This included all 11 participants who selected physician judgment at least once and all 7 who selected physician judgment in the salvage-treatment scenario.

Table 1. Participant Characteristics (N = 14)

Characteristic	n	%
Sex: Female	12	85.7
Sex: Male	2	14.3
Age, median (range), years	59.5 (40–85)	–
High-risk factor: None	11	78.6
High-risk factor: History of intraepithelial neoplasia or cancer	2	14.3
High-risk factor: HIV infection	1	7.1
Treatment experience: Radiotherapy	14	100.0
Treatment experience: Chemotherapy	14	100.0
Treatment experience: Local excision	1	7.1
Treatment experience: Abdominoperineal resection	1	7.1

Local excision and abdominoperineal resection were both experienced by the same participant.

Table 2. Responses Indicating Reliance on Physician Judgment by Clinical Scenario (N = 14 for each item)

Clinical scenario item	n	%
Diagnostic/evaluation-related items		
Factors important when determining timing of treatment-response assessment	0	0.0
Timing of treatment-response assessment	1	7.1
Screening in high-risk populations	2	14.3
Additional FDG-PET/CT for staging	2	14.3
Magnetic resonance imaging for response assessment	2	14.3
Treatment-related items		
Treatment goal in early anal cancer	3	21.4
Factors important when selecting chemotherapy	3	21.4
Treatment choice for stage I anal cancer	3	21.4
Low-dose vs standard-dose radiotherapy	4	28.6
Salvage treatment for suspected residual cancer	7	50.0

Completion of the two trade-off threshold items was also similar across these response patterns: 9 of 11 participants who selected physician judgment at least once and all 3 who never selected it completed both thresholds; among the 7 participants who relied on physician judgment for salvage treatment, 6 completed both thresholds, the same proportion as among the 7 who selected a substantive salvage-treatment option (Table 3). Thus, selection of physician judgment did not correspond to an absence of elicitable outcome-priority or trade-off information in this sample.

Cure was the highest-ranked treatment goal overall (mean rank, 2.00; median, 1; ranked first by 8 of 14 and in the top two by 10 of 14), followed by quality of life (mean rank, 2.71). Permanent stoma had a mean rank of 3.79, whereas time able to maintain usual activities (4.50) and treatment-related adverse effects (4.57) were ranked lowest on average (Supplementary Table S1). The three participants who never selected physician judgment all ranked cure first (Supplementary Table S2). By contrast, all six participants whose first-ranked goal was not cure selected physician judgment at least once; their mean number of physician-judgment selections was 3.00 compared with 1.12 among the eight participants who ranked cure first. Given the small groups, this pattern is considered exploratory.

General outcome priorities and context-specific trade-offs

The clearest within-participant example involved salvage treatment for suspected residual cancer. Seven participants (50.0%) responded that they would follow their treating

physician's judgment when asked to choose among abdominoperineal resection resulting in a permanent stoma, non-surgical treatment, or observation. Five participants (35.7%) preferred to avoid surgery because of the burden of a permanent stoma, whereas 2 (14.3%) would accept surgery if it offered the possibility of cure.

In the separate survival-threshold item, 13 participants indicated the improvement in survival they would require to accept surgery considering both the permanent stoma and operative risk. Ten selected an improvement of $\geq 30\%$, 1 selected 20%-30%, and 2 selected $\leq 5\%$. Among the 7 participants who had left the salvage-treatment decision to the physician, 4 selected $\geq 30\%$, 1 selected 20%-30%, and 2 selected $\leq 5\%$ (Table 4). Thus, 5 of 7 participants who relied on physician judgment for the treatment choice nevertheless indicated that a survival improvement of at least 20% would be required to accept the surgical burden. Across all respondents to the threshold item, 11 of 13 required at least a 20% survival improvement, even though permanent stoma ranked fourth among the six general treatment goals. This illustrates that general outcome rankings and context-specific trade-off thresholds provided complementary rather than interchangeable information.

A similar pattern of contextual preference was observed for stage I anal cancer. Twelve participants answered the additional-recurrence-risk threshold for choosing local excision rather than chemoradiotherapy. Among the 9 participants who selected chemoradiotherapy in the corresponding treatment-choice item, 8 indicated that they would accept no more than a 3% increase in local recurrence risk with local

Table 3. Completeness of Outcome-Preference Information According to Reliance on Physician Judgment

Response pattern	n	Complete six-goal ranking	Both threshold items completed
Physician judgment selected for salvage treatment	7	7/7 (100%)	6/7 (85.7%)
Substantive option selected for salvage treatment	7	7/7 (100%)	6/7 (85.7%)
Physician judgment selected at least once across 10 items	11	11/11 (100%)	9/11 (81.8%)
Physician judgment never selected across 10 items	3	3/3 (100%)	3/3 (100%)

The two threshold items assessed the maximum acceptable additional recurrence risk with local excision and the minimum survival improvement required to accept surgery resulting in a permanent stoma. "Substantive option" indicates selection of one of the treatment alternatives rather than physician judgment.

Table 4. Salvage-Treatment Choice and Survival-Improvement Threshold for Accepting Surgery (N = 14)

Response to salvage-treatment item	n	$\leq 5\%$	5-10%	11-20%	20-30%	$\geq 30\%$	Missing
Surgery if cure could be expected	2	0	0	0	0	2	0
Avoid surgery because of burden of permanent stoma	5	0	0	0	0	4	1
Follow treating physician's judgment	7	2	0	0	1	4	0
Total	14	2	0	0	1	10	1

Survival-improvement thresholds are presented using the mutually exclusive response categories in the original survey.

excision ([Supplementary Table S3](#)).

The two items that separately offered both “It is difficult to decide” and the physician-judgment option also suggested that these responses were not interchangeable. For the choice between low-dose and standard-dose radiotherapy, 4 participants selected physician judgment and none selected “It is difficult to decide.” For the timing of treatment-response assessment, one participant selected each response. Because only two items used this response structure, these observations are descriptive only.

Sensitivity analyses and response patterns

The screening-frequency item that combined “I do not know” with “I would follow my physician’s judgment” was excluded from the primary analysis because the two meanings could not be separated. If this combined category were nevertheless counted as physician reliance, the number of participants with at least one physician-reliance response would increase from 11 of 14 (78.6%) to 13 of 14 (92.9%) ([Supplementary Table S4](#)). This item therefore illustrates the measurement consequences of combining uncertainty and reliance on clinician judgment in a single response option.

One participant marked multiple non-physician options on three of the 10 single-choice items used in the physician-reliance analysis. Excluding this participant produced similar descriptive results: 7 of 65 diagnostic/evaluation responses (10.8%) and 19 of 65 treatment responses (29.2%) indicated physician reliance, and 10 of 13 participants (76.9%) selected physician judgment at least once.

The survival-threshold item showed substantial clustering in its highest category: 10 of 13 respondents selected $\geq 30\%$. Among the 12 participants who completed both threshold items, however, only one selected the final response category on both items, whereas four selected the first category for acceptable recurrence risk and the final category for required survival improvement ([Supplementary Table S5](#)). This pattern is not consistent with a uniform tendency to choose the same endpoint across questions, although the upper category of $\geq 30\%$ still created a ceiling that prevented more precise estimation of the required benefit for those respondents.

Participants who more frequently selected physician judgment also tended to select response options indicating no particular expected benefit, burden, or concern on some items; this descriptive pattern was driven largely by the magnetic-resonance-imaging burden item. Such responses could reflect weaker differentiation, lower engagement with some questions, or genuinely neutral perceptions. Importantly, participants with the highest frequencies of physician-judgment responses still completed the treatment-goal ranking

and provided distinct outcome-priority profiles, indicating that lower differentiation on some items should not be equated with an absence of outcome values.

Discussion

The principal finding of this exploratory study is that reliance on physician judgment and the absence of patient values and preferences should not be treated as equivalent. Most participants generally preferred joint decision-making, yet many relied on physician judgment in at least one specific scenario. More importantly, participants who selected physician judgment still provided complete rankings of treatment goals and, in most cases, explicit trade-off thresholds. These within-participant patterns show that a response about how a clinical decision should be made does not by itself determine whether a patient has preferences about the outcomes of that decision.

This interpretation is consistent with prior work on decision-making roles. Bradley et al. showed that a general preference for decisional control did not consistently predict preferences across specific clinical scenarios [5]. Flynn et al. further demonstrated that participation is multidimensional: preferences can differ across information exchange, discussion, and selection of the final treatment [4]. In cancer care, Keating et al. found that characteristics of the decision itself influenced whether patients reported patient-controlled, shared, or physician-controlled roles [6]. Tamirisa et al. likewise highlighted that patients may value shared decision-making while still seeking a physician recommendation [7]. The present study extends these observations in a guideline-development setting by linking scenario-specific reliance on physician judgment with independently elicited outcome priorities and quantitative treatment-acceptance thresholds within the same respondents.

The distinction is particularly important for clinical practice guideline methodology. In the GRADE Evidence to Decision framework, values refer to the importance patients place on the main outcomes of the options under consideration and to variability or uncertainty in those judgments [1,2]. A survey response about whether the patient would follow the physician’s judgment therefore does not directly measure how that patient values survival, treatment toxicity, quality of life, recurrence, or a permanent stoma. Coding such a response as “no preference,” undecided, or missing may consequently discard information that is directly relevant to judgments about the balance of desirable and undesirable effects and, ultimately, the direction or strength of a recommendation.

The present data further suggest that guideline-related preference studies should distinguish at least three analytically separable types of information. The first concerns the decision process: the extent to which the patient wishes to participate in deliberation and final choice or to rely on clinician judgment. The second concerns general outcome priorities: which consequences are regarded as more or less important when considered abstractly. The third concerns context-specific trade-offs: the magnitude of benefit or harm that changes whether a concrete option is acceptable. These should not be interpreted as psychometrically validated constructs on the basis of this small study, but they represent different forms of information that may answer different questions in guideline development.

The contrast between general rankings and contextual thresholds was especially informative. Permanent stoma was not among the highest-ranked goals when considered in the abstract, yet most respondents required a large survival improvement before accepting surgery that would result in a permanent stoma. Conversely, the recurrence-risk threshold for local excision was broadly consistent with the treatment choices of participants who preferred chemoradiotherapy. These findings do not imply that threshold questions are inherently superior to rankings. Rather, rankings identify relative priorities across outcomes, whereas thresholds reveal how those priorities operate when benefits and harms must be traded against one another. Both may therefore be useful, but they should not be treated as interchangeable measures of “patient values.”

An exploratory pattern also suggested that preserving physician-reliance responses may be particularly important for patients whose priorities are less dominated by cure. All participants whose first-ranked treatment goal was not cure selected physician judgment at least once, whereas the participants who never selected physician judgment all ranked cure first. The groups were too small for inference, and this observation requires replication. Nevertheless, it raises the possibility that treating physician-reliance responses as missing could disproportionately remove patients whose priorities differ from the most common cure-oriented pattern rather than remove only patients who lack preferences.

Reliance on physician judgment was observed more often in treatment-related than diagnostic/evaluation-related items in this sample. This may reflect the greater perceived consequences, uncertainty, complexity, or responsibility associated with treatment choices, but the present study cannot distinguish among these explanations. The diagnostic and treatment items were not matched for difficulty, information content, risk, response format, or burden. This find-

ing should therefore be regarded as a hypothesis-generating observation about context dependence rather than evidence that patients systematically prefer more physician control for treatment decisions.

The survey also illustrates a practical measurement problem. One item combined “I do not know” and “I would follow my physician’s judgment” into a single response category. Those responses cannot be classified as either uncertainty or reliance without making an assumption, and counting the combined category as physician reliance materially changed the headline proportion of participants with at least one such response. Guideline-related surveys should therefore avoid response options that merge lack of knowledge, decisional difficulty, neutrality, and reliance on clinician judgment, because these responses carry different meanings and may require different follow-up questions.

Strengths and limitations

A strength of this study is its within-participant design: general decision-making preference, responses across multiple real clinical scenarios, general treatment-goal rankings, and context-specific treatment-acceptance thresholds were linked within the same respondents. This allowed the analysis to address conceptual relationships that cannot be inferred from population-level proportions alone. The study also used patients with direct experience of anal cancer treatment recruited across four hospitals, providing an empirical example from a rare disease in which several guideline decisions involve substantial trade-offs.

The study also has important limitations. The sample was very small ($n = 14$) and recruited by convenience sampling, so the percentages cannot be generalized to the broader population of patients with anal cancer and the stability of the observed patterns is uncertain. The survey was developed for guideline development rather than as a validated instrument for the present analytic question, and response options differed across items. The forced ranking of six treatment goals demonstrates that outcome-priority information could be elicited, but it does not establish that respondents perceived the six goals as clearly or equally distinguishable. The threshold questions used categorical response ranges rather than a validated or iterative preference-elicitation method; in particular, the $\geq 30\%$ upper category for survival improvement produced a ceiling effect. General participation preference and scenario-specific reliance were not measured on the same scale, and only two items separately distinguished decisional difficulty from physician judgment.

The questionnaire was administered face-to-face by an outpatient-clinic nurse. This mode of administration could

have influenced some responses through social desirability or deference to healthcare professionals; however, the study did not assess the degree of prior rapport between respondents and the administering nurse, so the direction and magnitude of any such effect are uncertain. Some participants also gave less differentiated responses on selected items, and one participant marked multiple options on several single-choice questions, although sensitivity analysis excluding that participant did not materially change the principal frequencies.

Most respondents were women, all had received chemoradiotherapy, and only one had experienced a permanent stoma. Preferences based on anticipated outcomes and preferences informed by direct experience were therefore mixed. The survey did not ask why participants chose physician judgment—for example, trust in expertise, decision difficulty, responsibility burden, insufficient information, desire for recommendation, or another reason. Finally, because the analyses were exploratory and descriptive, the observed relationships among decision-process preferences, outcome priorities, and context-specific trade-offs should be treated as hypotheses for future study rather than as validated dimensions of a measurement model.

Future studies should deliberately separate preferences for information sharing, deliberation, and final decisional control from general outcome priorities and context-specific trade-offs. When participants choose to rely on physician judgment, follow-up questions should ask why and should continue to elicit the outcomes that matter and the benefit-risk trade-offs that would make an option acceptable. Preference-elicitation methods that allow more precise estimation beyond categorical ceiling values, combined with qualitative interviews, may be particularly useful. Given the rarity of anal cancer, larger multicenter studies will be needed to determine whether the patterns observed here are reproducible and how these different forms of preference information should be incorporated into guideline Evidence to Decision judgments.

Conclusion

In this exploratory multicenter analysis of patients with anal cancer, reliance on physician judgment in a specific clinical scenario did not necessarily indicate an absence of patient values or preferences. Patients who relied on physician judgment could still identify which outcomes mattered and specify the benefit-risk trade-offs that would make a treatment acceptable. For clinical practice guideline development, preference for the decision process, general outcome priorities,

and context-specific trade-offs should therefore be elicited and interpreted as distinct but related forms of information. Responses such as “I would follow my treating physician’s judgment” should not automatically be treated as no preference or missing data; rather, they should prompt further assessment of what outcomes matter and what trade-offs are acceptable.

Conflict of Interest

Hyun Jung Kim has been an editor of the Journal of Evidence-Based Practice since 2025. However, she was not involved in the peer reviewer selection, evaluation, or decision process of this article. No other potential conflicts of interest relevant to this article were reported.

Funding

The authors received no specific funding for this work.

Data Availability Statement

The datasets generated and/or analyzed in the current study are available from the corresponding author on reasonable request.

Ethics Approval and Consent to Participate

Not applicable.

Authors’ Contributions

Conceptualization: HJK, HSR, JMK. Data curation: HSR, JMK, KYL, DHK. Formal analysis: HJK, HSR, DHK, DK. Methodology: HJK, HS. Project administration: HSR, HS, JK. Visualization: JMK, DHK, HJK. Writing – original draft: HS, HJK, HSR, KYL. Writing – review & editing: HJK, JMK, DK, HS, HSR, KYL.

Acknowledgments

None.

ORCID

Hyo Seon Ryu, <https://orcid.org/0000-0003-2606-9973>

Jung-Myun Kwak, <https://orcid.org/0000-0002-2181-4279>

Kil-Yong Lee, <https://orcid.org/0000-0002-7949-0166>

Dalyong Kim, <https://orcid.org/0000-0003-0193-422X>

Dong Hyun Kang, <https://orcid.org/0000-0002-6559-0369>
Heather Swan, <https://orcid.org/0000-0001-9129-5008>
Hyun Jung Kim, <https://orcid.org/0000-0003-2018-2385>

Supplementary Materials

Supplementary materials can be found via <https://doi.org/10.63528/jebp.2026.00015>.

References

1. Alonso-Coello P, Oxman AD, Moher J, Brignardello-Petersen R, Akl EA, Davoli M, et al. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 2: clinical practice guidelines. *BMJ* 2016; 353: i2089.
2. Zhang Y, Alonso-Coello P, Brozek J, Wiercioch W, Etzean-dia-Ikobaltzeta I, Akl EA, et al. Using patient values and preferences to inform the importance of health outcomes in practice guideline development following the GRADE approach. *Health Qual Life Outcomes* 2017; 15: 52.
3. Elwyn G, Frosch D, Thomson R, Joseph-Williams N, Lloyd A, Kinnersley P, et al. Shared decision making: a model for clinical practice. *J Gen Intern Med* 2012; 27: 1361-7.
4. Flynn KE, Smith MA, Vanness D. A typology of preferences for participation in healthcare decision making. *Soc Sci Med* 2006; 63: 1158-69.
5. Bradley JG, Zia MJ, Hamilton N. Patient preferences for control in medical decision making: a scenario-based approach. *Fam Med* 1996; 28: 496-501.
6. Keating NL, Landrum MB, Arora NK, Malin JL, Ganz PA, van Ryn M, et al. Cancer patients' roles in treatment decisions: do characteristics of the decision influence roles? *J Clin Oncol* 2010; 28: 4364-70.
7. Tamirisa NP, Goodwin JS, Kandam A, Linder SK, Weller S, Turrubiate S, et al. Patient and physician views of shared decision making in cancer. *Health Expect* 2017; 20: 1248-53.
8. Stier EA, Clarke MA, Deshmukh AA, Wentzensen N, Liu Y, Poynten IM, et al. International Anal Neoplasia Society's consensus guidelines for anal cancer screening. *Int J Cancer* 2024; 154: 1694-702.
9. Feng M, Hallemeier CL, Almada C, Aranha O, Dorth J, Felder S, et al. Radiation therapy for anal squamous cell carcinoma: an ASTRO clinical practice guideline. *Pract Radiat Oncol* 2025; 15: 367-86.
10. Rao S, Guren MG, Khan K, Brown G, Renehan AG, Steigen SE, et al. Anal cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. *Ann Oncol* 2021; 32: 1087-100.

General anesthesia with remimazolam in a patient with Leigh syndrome: a case report

Young Ki Kim, Dong-Hyun Kang

Department of Anesthesiology and Pain Medicine, Gangneung Asan Hospital, College of Medicine, University of Ulsan, Gangwon, Korea

Background: Leigh syndrome is a mitochondrial disorder in which impaired oxidative phosphorylation causes progressive neurologic damage. Brainstem lesions may cause central hypoventilation, apnea, and recurrent pulmonary infection, increasing perioperative risk.

Case: We describe a patient with Leigh syndrome who had chronic central respiratory dysfunction, recurrent pneumonia, and a recent intensive care admission for sepsis and who subsequently underwent tracheostomy under general anesthesia. Anesthesia was induced and maintained with remimazolam and remifentanyl. Hemodynamic parameters remained acceptable neuromuscular recovery, spontaneous ventilation remained inadequate, and the patient was returned to the intensive care unit with ventilatory support.

Conclusion: In mitochondrial disease, propofol warrants careful consideration because of concerns regarding mitochondrial effects and propofol infusion syndrome. Remimazolam was selected based on severe respiratory vulnerability, recent multiorgan illness, and the anticipated need for postoperative ventilatory support, rather than as evidence of superiority over other agents.

Keywords: Leigh syndrome; remimazolam; central respiratory failure; mitochondrial disease; tracheostomy; general anesthesia

Leigh syndrome typically presents in infancy or childhood and results from impaired mitochondrial energy production. Neurologic deterioration is often accompanied by characteristic bilateral lesions in the basal ganglia, brainstem, and related structures, although the disorder is genetically and clinically heterogeneous [1,2].

Respiratory involvement is a major determinant of outcome in Leigh syndrome. Brainstem dysfunction may cause central apnea or hypoventilation, and bulbar dysfunction, ineffective cough, reduced mobility, and skeletal deformity can further impair airway clearance and increase the likelihood of recurrent pneumonia [2].

From an anesthetic perspective, the central issue is not only possible sensitivity to specific drugs but also the patient's limited physiologic reserve. Fasting, infection, temperature imbalance, hypoxemia, acid-base disturbance, and postoperative respiratory weakness may all contribute to

perioperative instability in patients with mitochondrial disease [3-5].

We report the use of remimazolam for tracheostomy in a young woman with Leigh syndrome complicated by chronic central respiratory failure, recurrent pneumonia, and a recent intensive care admission after sepsis.

Case Report

A 21-year-old woman with Leigh syndrome was scheduled for tracheostomy to secure a long-term airway because of recurrent pneumonia and respiratory failure requiring prolonged ventilatory support. General anesthesia was planned. She had severe psychomotor developmental delay, weighed 39.5 kg, and was 138 cm tall. Brain computed tomography showed marked cerebral and brainstem atrophy, necrotic changes in both basal ganglia, and diffuse ventriculomega-

Received: June 1, 2026; **Revised:** July 18, 2026; **Accepted:** July 30, 2026

Corresponding author: Young ki Kim

Email: ykkim@gnah.co.kr

© 2026 Korean Society of Evidence-Based Medicine

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

ly. She was bedridden and had scoliosis, diabetes mellitus, uncontrolled seizures, recurrent pneumonia, and recurrent pancreatitis. Two months before surgery, she developed septic shock due to multidrug-resistant *Acinetobacter baumannii*, which progressed to multiorgan failure. Eleven days before surgery, she underwent endotracheal intubation for acute respiratory acidosis. In view of recurrent pneumonia and repeated apneic episodes, tracheostomy was planned to provide a more stable long-term airway. Nutritional support was being delivered through a gastrostomy tube.

No premedication was given. On physical examination, the patient was drowsy and uncooperative. Heart sounds were normal, but coarse breath sounds were noted on auscultation. Chest anteroposterior radiography demonstrated severe thoracic scoliosis with rightward tracheal deviation of 3 cm. Preoperative laboratory findings were as follows: hemoglobin 9.0 g/dL, hematocrit 28.9%, blood glucose 138 mg/dL, and C-reactive protein 0.34 mg/dL. Because of apnea related to central respiratory failure and her recent septic condition, she was already receiving ventilatory support after intubation in synchronized intermittent mandatory ventilation-pressure support mode with an FiO₂ of 21%, PEEP of 6 cmH₂O, trigger flow of 5 L/min, pressure control level of 13 cmH₂O, pressure support level of 8 cmH₂O, and respiratory rate of 16 breaths/min. Arterial blood gas analysis showed a pH of 7.34, PaCO₂ of 48 mmHg, PaO₂ of 47 mmHg, base excess of -0.1 mmol/L, oxygen saturation of 79.7%, and lactate of 13 mg/dL.

These findings reflected the combined effect of central hypoventilation from Leigh syndrome and ongoing sepsis-related respiratory compromise. Despite mechanical ventilatory support, oxygenation remained inadequate, and tracheostomy was therefore indicated to secure a stable long-term airway rather than representing a decision to proceed with elective surgery in a compromised patient.

Because the patient had severe neurologic injury and marked respiratory vulnerability, general anesthesia with remimazolam and remifentanyl was planned in view of its hemodynamic stability and reversibility. Both induction and maintenance were achieved with remimazolam and remifentanyl. Remimazolam was administered as a loading infusion at 12 mg/kg/h for 2 minutes and then continued at 1 mg/kg/h for maintenance. Remifentanyl was infused continuously at 0.01 mcg/kg/min throughout the procedure. Rocuronium 10 mg was given for neuromuscular blockade. Before surgery, flumazenil, the antagonist of remimazolam, and sugammadex, the reversal agent for rocuronium, were prepared in case they were needed. The patient arrived in the operating room already intubated, and anesthesia was initiated through the existing endotracheal tube without any additional airway

device. On arrival, her heart rate was 83 beats/min and blood pressure was 150/90 mmHg. During surgery, the bispectral index was maintained between 40 and 50. Heart rate remained within 70-100 beats/min, and systolic blood pressure was stable between 130 and 160 mmHg. Hartmann solution was infused intraoperatively, with a total volume of 400 mL. Urine output was 60 mL, and body temperature remained stable at 37.0°C. The train-of-four count was 0 throughout the operation. The procedure lasted 605 seconds and was completed without complication. In the operating room, sugammadex 100 mg was administered, and the train-of-four recovered to 98%; however, adequate spontaneous ventilation did not return because of the patient's preexisting condition and central respiratory failure. She was therefore transferred to the intensive care unit and placed on pressure-controlled ventilation. One day later, spontaneous breathing gradually improved, and ventilatory support was changed to synchronized intermittent mandatory ventilation-pressure support mode with an FiO₂ of 21%, PEEP of 5 cmH₂O, trigger flow of 2 L/min, pressure control level of 12 cmH₂O, pressure support level of 12 cmH₂O, and respiratory rate of 17 breaths/min. Ten days after surgery, tracheal edema caused airway narrowing such that only one-eighth of the tracheal lumen was visible on endoscopy, and intravenous dexamethasone was administered for 3 days. Pleural effusion also increased, and she remained in the intensive care unit for observation for 4 weeks before transfer to the general ward. Even after transfer, sepsis and respiratory acidosis repeatedly worsened and improved. After a plan was made to replace the tracheostomy cannula every 1-2 months, she was discharged home with a home ventilator.

Discussion

Leigh syndrome is a neurometabolic disorder caused by mitochondrial dysfunction, and symmetric lesions in the brainstem and deep gray nuclei are among its characteristic radiologic and pathologic findings [1,2]. Because these regions are essential for respiratory control, airway protection, and autonomic regulation, anesthetic management is often challenged more by limited physiologic reserve than by any single uniform pharmacologic abnormality.

Respiratory failure is one of the major determinants of prognosis in Leigh syndrome [1]. Central hypoventilation, apneic episodes, ineffective cough, recurrent aspiration or infection, and reduced mobility may all contribute to prolonged ventilatory dependence. In the present case, repeated pneumonia, impaired consciousness, and prior intubation suggested a particularly high risk of postoperative respiratory

insufficiency.

Perioperative risk in Leigh syndrome should therefore be considered within the broader framework of mitochondrial disease and advanced neurologic impairment [3-5]. The interaction among chronic respiratory dysfunction, recent sepsis, nutritional stress, immobility, and impaired airway clearance may be clinically more important than any isolated concern regarding anesthetic drug sensitivity.

Accordingly, practical anesthetic goals in mitochondrial disease include minimizing prolonged fasting when possible, maintaining oxygenation and ventilation, preserving circulatory stability, maintaining normothermia, correcting metabolic disturbances, and planning carefully for postoperative respiratory support [3-5]. In patients who already have central respiratory failure, the postoperative strategy may be as important as the choice of hypnotic agent.

Propofol remains a widely used anesthetic agent, but caution is warranted in mitochondrial disorders because experimental studies have demonstrated inhibitory effects on mitochondrial function, including electron transport and fatty acid oxidation [6]. Clinical concern is reinforced by propofol-related infusion syndrome, a rare but potentially fatal complication that is usually associated with prolonged infusion, critical illness, or high cumulative exposure [7,8].

Propofol-related infusion syndrome has been linked to severe metabolic acidosis, rhabdomyolysis, hyperkalemia, cardiac dysfunction, renal failure, and death [8,9]. In addition, case reports suggest that unrecognized mitochondrial disease may increase susceptibility in some patients [9]. These findings do not make propofol absolutely contraindicated in all patients with mitochondrial disease, but they do support individualized assessment of dose, duration, and overall physiologic context [3,4].

Our decision-making therefore focused less on excluding a single drug categorically and more on selecting an anesthetic strategy that matched this patient's clinical condition. Given her recent multiorgan illness and substantial respiratory risk, we sought an approach that allowed controlled titration during surgery together with a clear postoperative airway plan.

Remimazolam is an ultra-short-acting benzodiazepine that is metabolized by tissue esterases and has been reported to provide predictable sedation or anesthesia with relatively stable hemodynamics [10-13]. Reversal with flumazenil is another practical advantage when rapid neurologic assessment is desired, although overall recovery still depends on the patient's underlying neurologic and respiratory status [12,14].

For these reasons, remimazolam was chosen in this case as an individualized option rather than as an established

best agent for Leigh syndrome. Our experience suggests that its use was feasible in this setting, but the finding should be interpreted in light of the patient's baseline risk for ventilator dependence, the limited value of a single case, and the ongoing uncertainty regarding anesthetic selection in mitochondrial disease.

The doses of remifentanyl and rocuronium were also selected with the patient's clinical condition in mind. Remifentanyl was initiated at a low infusion rate of 0.01 mcg/kg/min because of concern that higher doses might compromise hemodynamic stability in the setting of recent septic shock; because the bispectral index remained within the target range and blood pressure did not increase further during surgery, the rate was not titrated upward. Rocuronium was administered at a reduced dose of 10 mg (approximately 0.25 mg/kg) because progressive neuromuscular involvement in Leigh syndrome was expected to increase sensitivity to neuromuscular blocking agents; continuous train-of-four monitoring confirmed sustained relaxation throughout surgery without the need for additional doses.

In this patient, adequate spontaneous breathing did not return after recovery of the train-of-four at the end of surgery, and the need for transfer to the intensive care unit with ventilatory support is more plausibly explained by preexisting central respiratory failure and impaired brainstem function than by residual neuromuscular blockade [2,15]. Previous reports have also described difficult or delayed respiratory recovery after general anesthesia in patients with Leigh syndrome [1,2].

Moreover, the patient had already been intubated before surgery and required secure long-term airway access because of recurrent pneumonia and repeated apnea; immediate recovery of spontaneous breathing after tracheostomy was therefore not necessarily the primary postoperative goal. Continued ventilatory support in the intensive care unit, together with ongoing assessment of respiration, acid-base balance, infection, secretion clearance, and neurologic status, was more consistent with her underlying pathophysiology [3-5].

The clinical course and prognosis in this case cannot be attributed to a single anesthetic agent alone [4,5]. In addition to progressive neurologic injury from Leigh syndrome, the patient also had sepsis, recurrent pneumonia, scoliosis, poorly controlled seizures, diabetes mellitus, recurrent pancreatitis, and dependence on enteral feeding; the postoperative respiratory course and overall outcome were more likely shaped by multiple interacting factors [2,4,5].

Thus, although stable intraoperative hemodynamics were observed during anesthesia with remimazolam, this obser-

vation alone does not justify concluding that remimazolam improved the patient's long-term outcome or was superior to other anesthetic techniques [5,11,13,14,16].

This report has several limitations. First, as a single case, it cannot establish a causal relationship between anesthetic choice and clinical course, and its generalizability is limited. Second, because there was no direct comparison with other anesthetic strategies such as propofol, the relative advantages and disadvantages of remimazolam cannot be quantified [11,14,16]. Third, the patient had numerous confounding factors—including central respiratory failure, active infection, metabolic vulnerability, and structural thoracic abnormalities—making it difficult to isolate a single explanation for delayed postoperative respiratory recovery [2,15]. Fourth, long-term physiologic and clinical follow-up data, including mitochondrial function, lactate trends, and long-term outcome, were limited, which constrains interpretation of the longer-term implications of this anesthetic choice [4,5]. Ultimately, this case should be interpreted as a technical experience showing that remimazolam-based general anesthesia was feasible in a patient with severe Leigh syndrome, and further accumulation of cases and comparative studies is needed [5,14,16].

Conclusion

Leigh syndrome may confer substantial perioperative risk because mitochondrial dysfunction often coexists with brain-stem-related respiratory impairment, recurrent pulmonary complications, and limited physiologic reserve [2-4].

In this case, remimazolam and remifentanyl were used successfully for tracheostomy in the setting of recent sepsis, chronic central respiratory failure, and an anticipated need for postoperative ventilatory support.

This experience should not be interpreted as proof that remimazolam is superior to propofol or to other anesthetic techniques. Rather, it illustrates that anesthetic planning in mitochondrial disease should integrate drug selection with the patient's overall respiratory status, metabolic vulnerability, and postoperative airway strategy [3-5].

This case report was approved by the Institutional Review Board of Gangneung Asan Hospital (IRB No. 2026-04-002), which waived the requirement for informed consent due to the retrospective nature of the report. Written informed consent for publication could not be obtained as the patient's legal guardian was unreachable despite repeated attempts to make contact. The requirement for written informed consent was therefore formally waived by the Institutional Review Board of Gangneung Asan Hospital.

Conflict of Interest

The author declares no conflict of interest.

Funding

No funding was received for this work.

Data Availability Statement

Not applicable.

Ethics Approval and Consent to Participate

This case report was approved by the Institutional Review Board of Gangneung Asan Hospital (IRB No. 2026-04-002)

Authors' Contributions

Conceptualization: YKK, DHK. Methodology: YKK, DHK. Project administration: YKK, DHK. Writing – original draft: YKK, DHK. Writing – review & editing: YKK, DHK.

Acknowledgments

None.

ORCID

Young Ki Kim, <https://doi.org/0000-0002-7205-6746>

References

1. Bakare AB, Lesnefsky EJ, Iyer S. Leigh syndrome: a tale of two genomes. *Front Physiol* 2021; 12: 693734.
2. Sofou K, De Coo IF, Isohanni P, Ostergaard E, Naess K, De Meirleir L, et al. A multicenter study on Leigh syndrome: disease course and predictors of survival. *Orphanet J Rare Dis* 2014; 9: 52.
3. Niezgoda J, Morgan PG. Anesthetic considerations in patients with mitochondrial defects. *Paediatr Anaesth* 2013; 23: 785-93.
4. Terkawi AS, Wani TM, Al-Shuaibi KM, Tobias JD. Anesthetic considerations in Leigh disease: Case report and literature review. *Saudi J Anaesth* 2012; 6: 181-5.
5. Miyamoto Y, Miyashita T, Takaki S, Goto T. Perioperative considerations in adult mitochondrial disease: A case series and a review of 111 cases. *Mitochondrion* 2016; 26: 26-32.
6. Urban T, Waldauf P, Krajčová A, Jiroutková K, Halačková M, Džupa V. Kinetic characteristics of propofol-induced inhibition

- of electron-transfer chain and fatty acid oxidation in human and rodent skeletal and cardiac muscles. *PLoS One* 2019; 14: e0217254.
7. Hemphill S, McMenamin L, Bellamy MC, Hopkins PM. Propofol infusion syndrome: a structured literature review and analysis of published case reports. *Br J Anaesth* 2019; 122: 448-59.
 8. Krajčová A, Waldauf P, Anděl M, Duška F. Propofol infusion syndrome: a structured review of experimental studies and 153 published case reports. *Crit Care* 2015; 19: 398.
 9. Savard M, Dupré N, Turgeon AF, Desbiens R, Langevin S, Brunet D, et al. Propofol-related infusion syndrome heralding a mitochondrial disease: case report. *Neurology* 2013; 81: 770-1.
 10. Qiu Y, Gu W, Zhao M, Zhang Y, Wu J. The hemodynamic stability of remimazolam compared with propofol in patients undergoing endoscopic submucosal dissection: A randomized trial. *Front Med (Lausanne)* 2022; 9: 938940.
 11. Ko CC, Hung KC, Illias AM, Chiu CC, Yu CH, Lin CM, et al. The use of remimazolam versus propofol for induction and maintenance of general anesthesia: A systematic review and meta-analysis. *Front Pharmacol* 2023; 14: 1101728.
 12. Kilpatrick GJ. Remimazolam: non-clinical and clinical profile of a new sedative/anesthetic agent. *Front Pharmacol* 2021; 12: 690875.
 13. Liu X, Zhang L, Zhao L, Zhou X, Mao W, Chen L, et al. Comparison of the safety of remimazolam and propofol during general anesthesia in elderly patients: systematic review and meta-analysis. *Front Med (Lausanne)* 2025; 12: 1409495.
 14. Suzuki Y, Doi M, Nakajima Y. General anesthesia with remimazolam in a patient with mitochondrial encephalomyopathy: a case report. *JA Clin Rep* 2021; 7: 51.
 15. Cooper MA, Fox R. Anesthesia for corrective spinal surgery in a patient with Leigh's disease. *Anesth Analg* 2003; 97: 1539-41.
 16. Kuwabara T, Tamura T, Takakura M, Fujii T, Ozeki K, Akiyama K. Remimazolam-based anesthesia with intraoperative motor evoked potential monitoring in a patient with Leigh syndrome undergoing scoliosis surgery: a case report. *JA Clin Rep* 2026; 12:

Successful endoscopic internal drainage for infected intramural esophageal dissection complicated by mediastinal abscess: a case report

Jae Yong Park¹, Byung Joon Park², Beom Jin Kim¹, Jae Gyu Kim¹

¹Division of Gastroenterology, Department of Internal Medicine, Chung-Ang University College of Medicine, Seoul, Korea

²Department of Thoracic and Cardiovascular Surgery, Chung-Ang University Hospital, Chung-Ang University College of Medicine, Seoul, Korea

Background: Intramural esophageal dissection (IED) is a rare esophageal injury that is usually managed conservatively. However, the optimal treatment for complicated IED associated with infection or mediastinal abscess remains unclear.

Case: A 57-year-old man with decompensated liver cirrhosis presented with chest pain, odynophagia, and fever. Chest computed tomography revealed IED complicated by mediastinal abscess. Despite conservative management and thoracoscopic drainage, persistent fever suggested inadequate drainage of the infected false lumen. Endoscopic internal drainage was performed by longitudinal mucosal incision, followed by enteral nutritional support via feeding jejunostomy. The patient recovered without major surgery. Six-month follow-up endoscopy demonstrated complete healing of the false lumen with restoration of a single esophageal lumen.

Conclusions: Endoscopic internal drainage may be a feasible organ-preserving treatment option for selected patients with infected IED when adequate drainage cannot be achieved by conservative management or surgical drainage alone.

Keywords: Intramural esophageal dissection; False lumen; Mediastinal abscess; Endoscopic internal drainage; Liver cirrhosis; Case report

Introduction

Intramural esophageal dissection (IED) is a rare esophageal injury characterized by separation of the mucosal and sub-mucosal layers, resulting in the formation of a false lumen [1,2]. Although most cases are managed conservatively and have a favorable prognosis, complicated cases associated with infection or mediastinal abscess are extremely uncommon, and their optimal management remains unclear [2-4]. We report a case of infected IED complicated by mediastinal abscess that was successfully treated with endoscopic internal drainage following surgical drainage while avoiding esophagectomy.

Case Report

A 57-year-old man with alcohol-related liver cirrhosis complicated by esophageal varices, paroxysmal atrial fibrillation receiving direct oral anticoagulants, diabetes, and hypertension presented to the emergency department with worsening pain extending from the neck to the retrosternal area, odynophagia, and fever. On admission, his blood pressure was 127/75 mmHg, heart rate 81 beats/min, respiratory rate 20 breaths/min, and temperature 38.2°C. Laboratory findings showed a white blood cell count of 5,300/ μ L (74.0% neutrophils), hemoglobin of 12.9 g/dL, platelet count of 62×10^3 / μ L, CRP of 87.5 mg/L, total bilirubin of 1.7 mg/dL, albumin of

Received: August 11, 2026; **Received:** August 25, 2026; **Accepted:** August 30

Corresponding author: Jae Gyu Kim

Email: jgkimd@cau.ac.kr

© 2026 Korean Society of Evidence-Based Medicine

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

3.1 g/dL, AST/ALT of 30/18 IU/L, and PT-INR of 1.51. Contrast-enhanced chest computed tomography (CT) revealed a linear intramural defect in the mid-thoracic esophagus with diffuse submucosal fluid collection, suggestive of intramural esophageal dissection (IED), accompanied by superior mediastinal abscess formation with extraluminal air, suggesting possible microperforation (Fig. 1 A, B). Small bilateral pleural effusions were also present.

The patient was managed conservatively with bowel rest and intravenous ceftriaxone plus metronidazole. However, fever and chest pain persisted despite treatment. On hospital day 5, video-assisted thoracoscopic surgery was performed. A 1-cm abscess pocket above the azygos vein was identified and drained, and a right chest tube was inserted. No gross esophageal perforation or fistulous communication was identified. Upper endoscopy performed on hospital day 6

demonstrated diffuse edematous changes with multiple mucosal openings communicating with the intramural false lumen, from which blood-tinged purulent fluid was observed draining (Fig. 2 A, B). Grade I esophageal varices were also present.

Although inflammatory markers gradually decreased after surgery, persistent fever raised concern for inadequate drainage of the infected intramural cavity. Therefore, therapeutic endoscopic intervention was planned on hospital day 10. Under general anesthesia, the largest mucosal opening was longitudinally extended from 28 to 35 cm from the upper incisors using an Insulated Tip knife to facilitate internal drainage (Fig. 3 A-D). Only minor transient oozing occurred during the procedure, without significant bleeding. Esophageal varices remained stable throughout the procedure. Further incision was not attempted because of marked mural

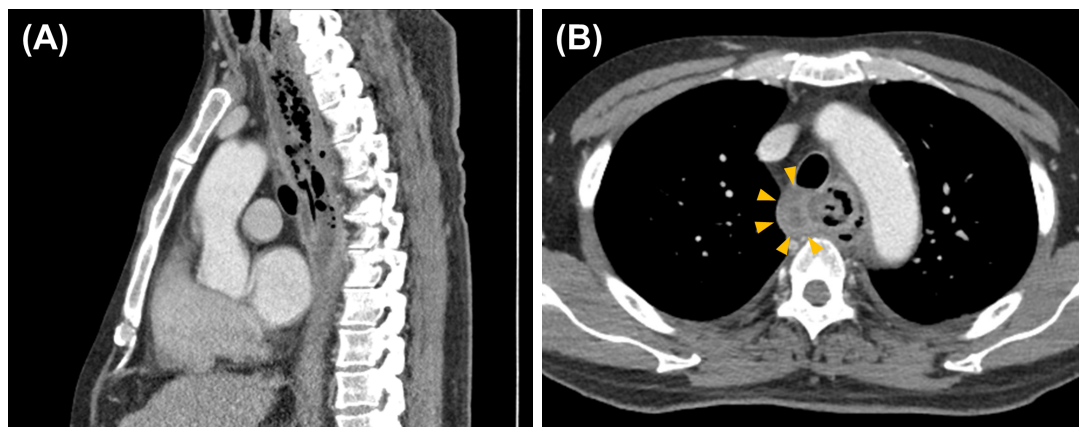


Fig. 1. Chest CT demonstrating (A) intramural esophageal dissection with (B) mediastinal abscess formation (arrowheads).

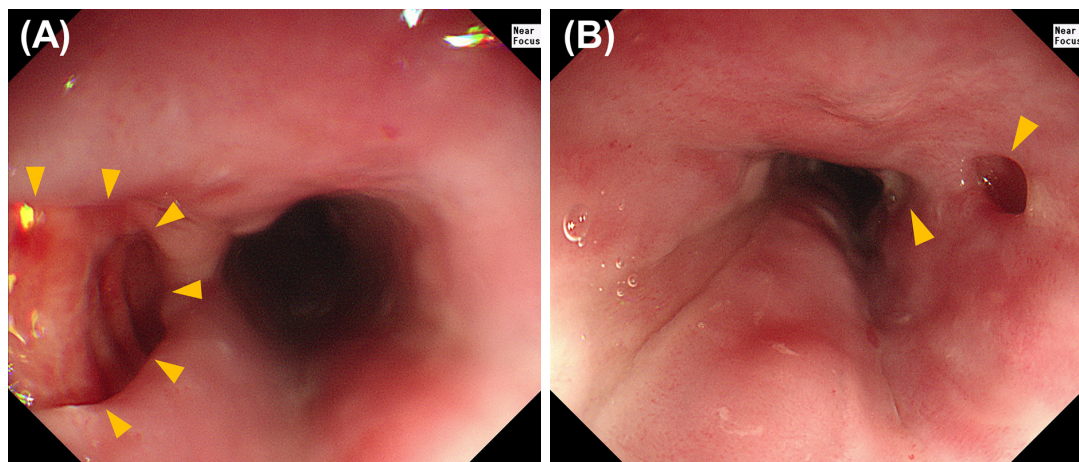


Fig. 2. Initial endoscopic findings. Upper endoscopy performed on hospital day 6 demonstrated diffuse edematous changes throughout the esophagus. (A) A large proximal mucosal opening (arrowheads; approximately 1.5 cm in diameter) located 28 cm from the upper incisors, communicating with a longitudinal intramural cavity (false lumen). (B) Multiple distal spontaneous mucosal openings with blood-tinged purulent drainage (arrowheads).

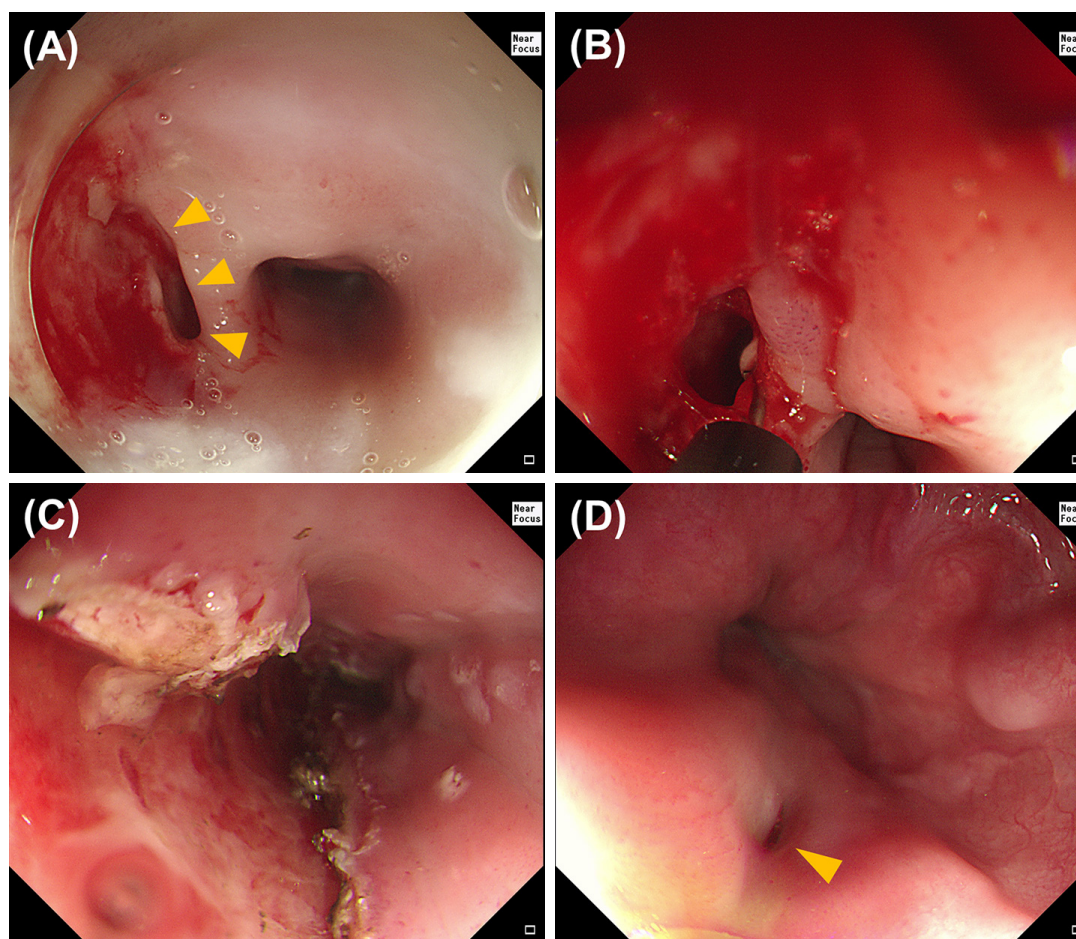


Fig. 3. Endoscopic internal drainage by longitudinal mucosal incision. (A) A large proximal mucosal opening at 28–29 cm from the upper incisors (9 o'clock position; arrowheads). (B) Longitudinal extension of the mucosal opening to approximately 35 cm using an Insulated Tip Knife to facilitate internal drainage. (C) Post-procedural appearance showing the widely opened false lumen with (D) persistent spontaneous drainage through multiple distal mucosal openings (arrowhead). No significant procedure-related bleeding was observed.

edema, luminal narrowing, underlying coagulopathy, the presence of esophageal varices, and evidence of ongoing spontaneous drainage of purulent fluid through the distal mucosal openings. Because prolonged fasting was anticipated, a radiologically guided feeding jejunostomy was placed on hospital day 12, and enteral nutrition was initiated. After 17 days of intravenous antibiotics, the inflammatory markers markedly improved. Following chest tube removal, the patient was discharged with amoxicillin/clavulanate administered via the feeding jejunostomy.

Serial chest CT examinations performed at 1- to 2-month intervals continued to suggest residual communication between the true and false lumens in the upper thoracic esophagus. During follow-up, the patient prematurely resumed oral intake against medical advice, resulting in transient mild aggravation of the radiologic findings. Therefore, conservative management with enteral nutrition and antibiotics was

continued. After normalization of inflammatory markers and sustained clinical improvement, follow-up endoscopy was performed 6 months after the initial presentation. Endoscopy demonstrated complete epithelialization of the previous false lumen with restoration of a single patent esophageal lumen (Fig. 4). Mild residual narrowing remained but did not impede endoscope passage. Oral feeding was resumed gradually without recurrence of symptoms.

Discussion

IED is a rare esophageal injury characterized by separation of the mucosal and submucosal layers, resulting in the formation of a false lumen [1,2]. It has been associated with esophageal instrumentation, forceful vomiting, coagulopathy, anticoagulant therapy, and underlying esophageal disorders [1,5]. Patients typically present with acute chest pain, odyno-

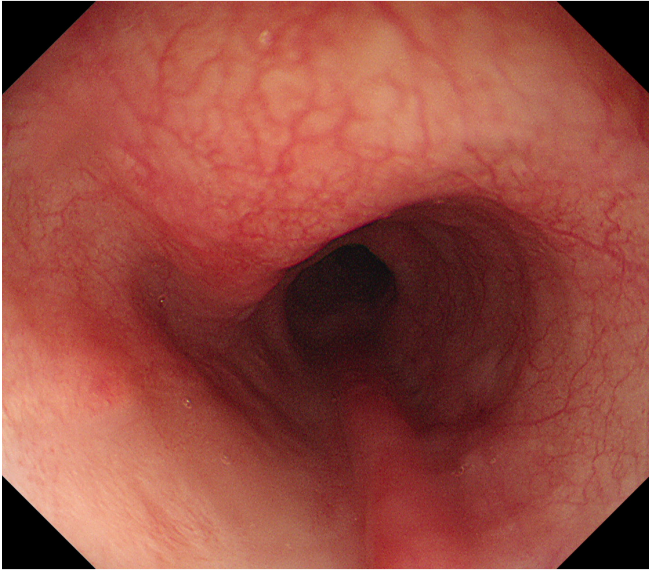


Fig. 4. Follow-up endoscopy 6 months after treatment showing complete epithelialization of the false lumen with restoration of a single esophageal lumen.

phagia, or dysphagia, which may mimic more common cardiopulmonary conditions [1,2]. Contrast-enhanced CT and upper endoscopy play complementary roles in establishing the diagnosis by demonstrating the false lumen and excluding transmural perforation [2,4].

Most reported cases of IED follow a benign clinical course and respond well to conservative management consisting of fasting, nutritional support, acid suppression, and antibiotics when infection is suspected [1,3]. In contrast, complicated IED accompanied by secondary infection or mediastinal abscess is exceedingly uncommon, and the optimal management strategy has not been established [4,5]. Although surgical intervention may be required in patients with uncontrolled sepsis or suspected perforation, esophagectomy is a highly invasive procedure associated with substantial morbidity and mortality, particularly in patients with advanced liver cirrhosis or other significant comorbidities [6,7]. Therefore, organ-preserving treatment strategies are desirable whenever clinically feasible.

Our patient presented with infected IED, as evidenced by persistent fever, elevated inflammatory markers, purulent drainage from the false lumen, and concomitant mediastinal abscess formation. Despite thoracoscopic drainage of the mediastinal abscess, persistent fever suggested inadequate drainage of the infected false lumen. Endoscopic longitudinal mucosal incision successfully converted the narrow mucosal opening into a wider internal drainage tract, facilitating continuous drainage of the infected cavity [8,9]. Clinical and

radiologic resolution was subsequently achieved without additional major esophageal surgery, suggesting that endoscopic internal drainage may serve as an effective adjunctive treatment when surgical drainage alone is insufficient [8-10].

In this case, decompensated liver cirrhosis complicated by esophageal varices, recurrent ascites, and hepatic encephalopathy was considered to confer a substantial perioperative risk for esophagectomy [6,7]. In this context, an organ-preserving approach was particularly desirable. To minimize the risk of bleeding, the mucosal incision was intentionally limited to the proximal lesion. Further distal extension was avoided because spontaneous drainage through multiple distal mucosal openings was already evident, while marked mural edema, luminal narrowing, and the presence of esophageal varices increased the anticipated procedural risk. The procedure was completed without clinically significant bleeding, and follow-up endoscopy demonstrated complete healing of the false lumen with restoration of a single esophageal lumen.

In conclusion, endoscopic internal drainage may represent a feasible organ-preserving treatment option for selected patients with infected IED, particularly when extensive surgical resection carries substantial operative risk [6-10]. Further studies are warranted to better define its role in the management of complicated IED.

Conflict of Interest

The authors declare no conflict of interest.

Funding

No funding was received for this work.

Data Availability Statement

Not applicable.

Ethics Approval and Consent to Participate

Institutional Review Board exemption was granted by the Institutional Review Board of Chung-Ang University Hospital (IRB No. 2608-001-19639), because written informed consent for publication could not be obtained due to loss to follow-up.

Authors' Contributions

Conceptualization: JYP. Patient management and investi-

gation: JYP and BJP. Writing – original draft: JYP. Writing – review & editing: JYP, BJP, BJK, and JGK. All authors read and approved the final manuscript.

Acknowledgments

None.

ORCID

Jae Yong Park, <https://orcid.org/0000-0001-6114-8920>

Byung Joon Park, <https://orcid.org/0000-0001-8746-3064>

Beom Jin Kim, <https://orcid.org/0000-0002-4841-9404>

Jae Gyu Kim, <https://orcid.org/0000-0002-0938-6697>

References

1. Beumer JD, Devitt PG, Thompson SK. Intramural oesophageal dissection. *ANZ J Surg* 2010; 80: 91-5.
2. Phan GQ, Heitmiller RF. Intramural esophageal dissection. *Ann Thorac Surg* 1997; 63: 1785-6.
3. Cheng G, Wang L, You J, Cheng L, Yang F, Wang X, et al. Asymptomatic intramural esophageal dissection: A case report and literature review. *Medicine (Baltimore)* 2025; 104: e42632.
4. Kumar S, Sakthivel MK, Bosemani T. Intramural Esophageal

Abscess Complicated with Pleural Fistula: A Case Report. *Cureus* 2020; 12: e6846.

5. Bak YT, Kwon OS, Yeon JE, Kim JS, Byun KS, Kim JH, et al. Endoscopic treatment in a case with extensive spontaneous intramural dissection of the oesophagus. *Eur J Gastroenterol Hepatol* 1998; 10: 969-72.
6. Low DE, Kuppusamy MK, Alderson D, Cecconello I, Chang AC, Darling G, et al. Benchmarking Complications Associated with Esophagectomy. *Ann Surg* 2019; 269: 291-8.
7. Sozzi M, Siboni S, Asti E, Bonitta G, Bonavina L. Short-Term Outcomes of Minimally Invasive Esophagectomy for Carcinoma In Patients with Liver Cirrhosis. *J Laparoendosc Adv Surg Tech A* 2017; 27: 592-6.
8. Cho CM, Ha SS, Tak WY, Kweon YO, Kim SK, Choi YH, et al. Endoscopic incision of a septum in a case of spontaneous intramural dissection of the esophagus. *J Clin Gastroenterol* 2002; 35: 387-90.
9. Zhang J, Wang H, Liu D. Treatment of a spontaneous intramural esophageal abscess by endoscopic mucosal incision: an easy and effective therapy. *Endoscopy* 2022; 54: e937-8.
10. Benatta MA, Grimaud JC, Kaci M, Desjeux A, Baghdadi M, Loumi A. Intramural esophageal dissection due to pharyngeal abscess treated by endoscopic esophageal transection: a case report. *Gastroenterol Clin Biol* 2010; 34: 329-31.

Instructions for authors

Journal of Evidence-Based Practice (J Evid-Based Pract) is the official scientific journal of the Korean Society of Evidence-Based Medicine. The abbreviated title is “*J Evid-Based Pract*”. It is published in English two times a year on the last day of March and September.

I. Aims and Scope

J Evid-Based Pract aims to present 1) Original evidence-based research on important issues in healthcare, 2) Methods, tools, and concepts essential for evidence-based medicine (EBM), education and practice, 3) Perspectives, debates, analyses, and opinions on reliable evidence and related topics in evidence-based medicine.

II. Editorial Policy

The Editor assumes that all authors listed in a manuscript have agreed with the following policy of the *J Evid-Based Pract* on submission of manuscript. Except for the negotiated secondary publication, the manuscript submitted to the *J Evid-Based Pract* must be previously unpublished and not be under consideration for publication elsewhere. Under any circumstances, the identities of the referees will not be revealed. All published manuscripts become the permanent property of the Korean Society of Evidence-Based Medicine (KSEBM) and may not be published elsewhere without written permission. *J Evid-Based Pract* adheres completely to guidelines and best practices published by professional organizations, including Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals (<http://www.icmje.org/icmje-recommendations.pdf>) from ICMJE and Principles of Transparency and Best Practice in Scholarly Publishing (joint statement by COPE, DOAJ, WAME, and OASPA; <http://doaj.org/bestpractice>) if otherwise not described below.

III. General information

1. Publication types

Manuscripts submitted to *J Evid-Based Pract* should present evidence-based research on important healthcare issues or contribute to the education and advancement of evidence-based medicine (EBM). Submissions must be unique, creative, and contribute meaningfully to the field. The jour-

nal accepts various types of manuscripts, including editorials, original articles, reviews, systematic review, clinical trial, clinical practice guideline, case reports, and letters to the editor.

2. Language

J Evid-Based Pract publishes articles in English. Spellings should abide by American spellings. Medical terminology should be written based on the most recent edition of Dorland's Illustrated Medical Dictionary. Accepted manuscripts are requested to be proofread by professional English editors.

3. Submission of manuscripts

In addition to members of Korean Society of Evidence-Based Medicine, any researcher throughout the world can submit a manuscript if the scope of the manuscript is appropriate. Authors are requested to submit their papers to ksebm.office@gmail.com via e-mail. Final revisions by authors should be submitted within 1 week of the request.

4. Data Availability Statement

Data sharing is encouraged by the *J Evid-Based Pract*, but a Data Availability Statement will be required and published with the manuscript. Authors will be provided the following options during submission or may use a draft of their own.

- The datasets generated during and/or analyzed during the current study are available in the [NAME] repository, [PERSISTENT WEB LINK TO DATASETS]
- The datasets generated during and/or analyzed during the current study are not publicly available due [REASON WHY DATA ARE NOT PUBLIC] but are available from the corresponding author on reasonable request.
- The datasets generated during and/or analyzed during the current study are available from the corresponding author upon reasonable request.
- Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.
- All data generated or analyzed during this study are included in this published article [and its supplementary information files].
- The data that support the findings of this study are available from [third party name] but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable re-

quest and with permission of [third party name].

5. Preprint policy

A preprint can be defined as a version of a scholarly paper that precedes formal peer review and publication in a peer-reviewed scholarly journal. *J Evid-Based Pract* allows authors to submit a manuscript that have been posted on preprint platform to the journal. It is not treated as duplicate submission or duplicate publication. *J Evid-Based Pract* recommend authors to disclose it with only single DOI during the submission process. Otherwise, it may be screened from the plagiarism check program — Similarity Check (iThenticate).

Preprint submission will be processed through the usual peer-review process. In addition, the preprint's history will be tracked by additional independent editor, with an emphasis on the posting procedure and format.

If the manuscript with preprint is accepted for publication, authors are recommended to update the information at the preprint platform with a link to the published article in *J Evid-Based Pract*, including DOI at *J Evid-Based Pract*. It is strongly recommended that authors cite the article in *J Evid-Based Pract* instead of the preprint.

Moreover, *J Evid-Based Pract* does not permit referencing a preprint as a reference unless there is an exceptional circumstance that the authors can justify.

If the authors of a submitted article differ from those of the preprint, the authors must explain the change in authorship and demonstrate that it complies with ICMJE recommendations.

6. Disclosure of Artificial Intelligence (AI) Programs

Artificial Intelligence (AI) programs (e.g. ChatGPT or other similar software) cannot be considered as authors of submitted manuscripts because they do not meet the requirements for authorship. For instance, they cannot understand the role of authors or take responsibility for the content of the paper. Additionally, AI cannot meet the authorship criteria set by organizations such as the International Committee of Medical Journal Editors (ICMJE). This includes having the ability to give final approval for publication and being accountable for the accuracy and integrity of the work.

Furthermore, AI lacks the capacity to comprehend a conflict of interest statement, and cannot legally sign such a statement. Additionally, AI does not have independent affiliation from its creators, nor can it hold copyright.

Therefore, when submitting a paper, authors should not include AI as authors but rather acknowledge the use of AI

and provide transparent information about how it was used in writing the manuscript. As the field of AI is rapidly evolving, authors using AI should declare this fact and provide specific technical details about the AI model used, including its name, version, source, and the method of application in the paper. This is in line with the ICMJE recommendation of acknowledging writing assistance.

7. Peer review process

- The *J Evid-Based Pract* received the papers via ksebm.office@gmail.com.
- Manuscripts to be reviewed: All submitted manuscripts are peer reviewed. Commissioned manuscripts are also reviewed. Research data or supplementary materials are subjected to peer review.
- Who conducts peer review: Submitted manuscripts will be reviewed by 2 or more external experts in the corresponding field. The editor selects peer reviewers according to the recommendation of the Editorial Board members or from the external expert database maintained by the editorial office. Some publication types, including editorials, errata, corrigenda, retraction, withdrawal, and letters to the editor, are reviewed by the editorial board member without external peer review.
- Type of peer review: *J Evid-Based Pract* uses double-blind review, which means that both the reviewer's and author's identities are concealed from the reviewers, and vice versa, throughout the review process. To facilitate this anonymous review, authors need to ensure that their manuscripts are prepared in a way that does not give away their identity. The names of reviewers are not posted in the published article.
- Screening before peer review: The manuscript is first reviewed for its format and adherence to the aims and scope of the journal. If the manuscript does not align with the aims and scope of the Journal or does not adhere to the Instructions for authors, it may be returned to the author immediately after receipt and without a review.
- Duration for the first decision: The result of the first peer review is usually finished within two months. If there is no correspondence from the editorial office on the fate of the submitted manuscript two months after the submission, please get in touch with the editorial office via ksebm.office@gmail.com
- Revision process: The Editorial Board may request authors to revise the manuscripts according to the reviewer's opinion. After revising the manuscript, the author should send the revised files with a reply to each item of the reviewer's opinion. Additions and amendments to the revised manu-

script should be highlighted in red. The author's revisions should be completed within 60 days after the request. If it is not received by the due date, the Editorial Board will not consider it for publication. To extend the revision period to more than 60 days, the author should negotiate with the Editorial Board. The manuscript review process should be finished with the second review. If the reviewers wish further review, the Editorial Board may consider it. Statistical editing is also performed if data need professional statistical review by a statistician. *J Evid-Based Pract* neither guarantees acceptance without review nor very short peer review times for unsolicited manuscripts.

- Final decision maker: The Editorial Board will make a final decision on the approval for publication of the submitted manuscripts and can request any further corrections, revisions, and deletions of the article text if necessary.
- The publication date is published with all published papers, including dates of submission, revision, and acceptance.
- Review of in-house manuscripts: All manuscripts from editors, staff, or editorial board members are subject to the same review process as other submissions. During the review process, they will not be involved in the selection of reviewers or the decision-making process. Editors will not handle their manuscripts even if they have been commissioned. The review and publication processes not described in the Instructions for Authors will be incorporated into the Editorial Policy Statements approved by the Council of Science Editor Board of Directors, available at <http://www.councilscienceeditors.org>.

8. Article processing charge and publication fee

J Evid-Based Pract has no author submission fees or other publication-related charges. All publication costs are supported by the publisher. *J Evid-Based Pract* is a platinum open access journal that does not charge author fees.

9. Copyrights and secondary publication

The *J Evid-Based Pract* owns copyrights of all published materials. On behalf of the co-author(s), the corresponding author must complete and submit the journal's copyright transfer agreement, which includes a section on the disclosure of potential conflicts of interest based on the recommendations of the International Committee of Medical Journal Editors, "Uniform Requirements for Manuscripts Submitted to Biomedical Journals". A copy of the form is made available to the submitting author within the online manuscript submission process. It is possible to republish manuscripts if ONLY the manuscripts satisfy the condition of secondary

publication of the Uniform Requirements for Manuscripts Submitted to Biomedical Journals, available at: <http://www.icmje.org>

10. Open access

J Evid-Based Pract is an Open Access journal accessible for free on the Internet. Accepted peer-reviewed articles are freely available on the journal website for any user, worldwide, immediately upon publication without additional charge.

IV. Research and Publication Ethics Guidelines

For the policies on research and publication ethics, the "Good Publication Practice Guidelines for Medical Journals" (https://www.kamje.or.kr/board/view?b_name=bo_publication&bo_id=13) or the "Ethical Guidelines on Good Publication" (<http://publicationethics.org/resources/guidelines>) or "Ethical Considerations in the International Committee of Medical Journal Editors" (<http://www.icmje.org/recommendations>) are applied.

1. Conflict-of-interest statement

The corresponding author is required to summarize all authors' conflict of interest disclosures. The disclosure form shall be same with ICMJE Uniform Disclosure Form for Potential Conflicts of Interest (www.icmje.org/conflicts-of-interest). A conflict of interest may exist when an author (or the author's institution or employer) has financial or personal relationships or affiliations that could influence (or bias) the author's decisions, work, or manuscript. All authors should disclose their conflicts of interest, i.e., (1) financial relationships (such as employment, consultancies, stock ownership, honoraria, paid expert testimony), (2) personal relationships, (3) academic competition, and (4) intellectual passion. These conflicts of interest must be included as a footnote on the title page or in the Acknowledgements section.

All funding sources should be declared on the title page or in the Acknowledgements section at the end of the text. If an author's disclosure of potential conflicts of interest is determined to be inaccurate or incomplete after publication, a correction will be published to rectify the originally published disclosure statement, and additional action may be taken as necessary.

If one or more editors are involved as authors, the authors should declare conflict of interest.

Ex) AAA has been an editor of the Journal of Evidence-Based Practice since 2017; however, he was not involved in the

peer reviewer selection, evaluation, or decision process of this article. No other potential conflicts of interest relevant to this article were reported.

2. Statement of informed consent

Copies of written informed consents and Institutional Review Board (IRB) approval for clinical research are recommended to be kept. The editor or reviewers may request copies of these documents to clarify potential ethical issues.

3. Protection of privacy, confidentiality, and written informed consent

Identifying details should not be published in written descriptions, photographs, or pedigrees unless it is essential for scientific purposes and the patient (or his/her parents or guardian) provides written informed consent for publication. Additionally, informed consent should be obtained in the event that the anonymity of the patient is not assured. For example, masking the eye region of patients in photographs is not adequate to ensure anonymity. If identifying characteristics are changed to protect anonymity, authors should assure that alterations do not distort scientific meaning. When informed consent has been obtained, this should be indicated in the published article.

4. Protection of human and animal rights

In the reporting of experiments that involve human subjects, it should be stated that the study was performed according to the Helsinki Declaration of 1975 (revised 2013) (Available from <https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/>) and approved by the Institutional Review Board (IRB) of the institution where the experiment was performed. Clinical studies that do not meet the Helsinki Declaration will not be considered for publication. Identifying details should not be published (such as name, initial of name, ID numbers, or date of birth).

In the case of an animal study, a statement should be provided indicating that the experimental processes, such as the breeding and the use of laboratory animals, were approved by the Research Ethics Committee (REC) of the institution where the experiment was performed or that they did not violate the rules of the REC of the institution or the NIH Guide for the Care and Use of Laboratory Animals (Institute of Laboratory Animal Resources, Commission on Life Sciences, National Research Council, <https://www.nap.edu/catalog/5140/guide-for-the-care-and-use-of-laboratory-animals>). The authors should preserve raw experimental study data for

at least 1 year after the publication of the paper and should present this data if required by the Editorial Board.

5. Registration of the clinical research

All prospective studies must be registered in the primary registry before submission. *J Evid-Based Pract* accepts registration in any of the primary registries that participate in the World Health Organization (WHO) International Clinical Trials Portal (<http://www.who.int/ictrp/en>), NIH ClinicalTrials.gov (<http://www.clinicaltrials.gov>), or Korea Clinical Research Information Service (CRIS, <http://cris.nih.go.kr>).

6. Reporting guidelines

The *J Evid-Based Pract* recommends that a submitted manuscript follow reporting guidelines appropriate for various study types. Good sources for reporting guidelines are the EQUATOR Network (www.equatornetwork.org) and the NLM's Research Reporting Guidelines and Initiatives (www.nlm.nih.gov/services/research_report_guide.html).

7. Author and authorship

An author is considered to be an individual who has made substantive intellectual contributions to a published study and whose authorship continues to have important academic, social, and financial implications.

Authorship credit should be based on: (1) substantial contributions to the conception or design of the work, or to the acquisition, analysis, or interpretation of data for the work; (2) the drafting of the article or revising it critically for important intellectual content; (3) final approval of the version to be published; and (4) agreement on taking accountability for the accuracy or integrity of the work. Authors should meet these four criteria, and these criteria distinguish the authors from other contributors.

Correction of authorship after publication: *J Evid-Based Pract* does not correct authorship after publication unless a mistake has been made by the editorial staff. Authorship may be changed before publication but after submission when an authorship correction is requested by all of the authors involved with the manuscript.

When a large, multicenter group has conducted the work, the group should identify the individuals who accept direct responsibility for the manuscript. When submitting a manuscript authored by a group, the corresponding author should clearly indicate the preferred citation and identify all individual authors as well as the group name. Acquisition of funding, collection of data, or general supervision of the research group alone does not constitute authorship. Journals gener-

ally list other members of the group in the Acknowledgments section.

8. Plagiarism and duplicate publication

Plagiarism is the use of previously published material without attribution. Prior to peer review, all manuscripts are screened for plagiarism by the Editor-in-Chief using iThenticate. When plagiarism is detected at any time before publication, the *J Evid-Based Pract* editorial office will take appropriate action as directed by the standards set forth by the Committee on Publication Ethics (COPE). For additional information, please visit <http://www.publicationethics.org>. Text copied from previously published work is interpreted using the following taxonomy:

1) Intellectual theft

Deliberate copying of large blocks of text without attribution

2) Intellectual sloth

Copying of “generic” text, e.g., a description of a standard technique, without clear attribution

3) Plagiarism for scientific English

Copying of verbatim text, often from multiple sources

4) Technical plagiarism

Use of verbatim text without identifying it as a direct quotation but citing the source

5) Self-“plagiarism”

Manuscripts are only accepted for publication if they have not been published elsewhere. Manuscripts published in this journal should not be submitted for publication elsewhere. Duplicate submissions identified during peer review will be immediately rejected, and duplicate submissions that are discovered after publication will be retracted. It is mandatory for all authors to resolve any copyright issues when citing a figure or table from a different journal that is not open access.

When a duplicate publication is detected, the *J Evid-Based Pract* editorial office will notify the counterpart journal of this violation. Additionally, it will be notified of the authors’ affiliation, and penalties will be imposed on the authors. It is possible to republish manuscripts if they satisfy the condition of secondary publication of the Uniform Requirements for Manuscripts Submitted to Biomedical Journals, available at: www.icmje.org. If the author or authors wish to obtain a duplicate or secondary publication for reasons such as publication for readers of a different language, the author(s) should obtain approval from the Editors-in-Chief of both the first and second journal.

V. Manuscript Preparation

J Evid-Based Pract recommends compliance with some or all of the following guidelines (<https://www.equator-network.org>).

CONSORT for reporting of randomized controlled trials (<http://www.consort-statement.org>)

STARD for reporting of diagnostic accuracy studies (<http://www.stard-statement.org>)

STROBE for reporting of observational studies in epidemiology (<http://www.strobe-statement.org>)

PRISMA for reporting of systematic reviews (<http://www.prisma-statement.org>)

MOOSE for reporting of Meta-analyses of observational studies (<https://jamanetwork.com/journals/jamasurgery/article-abstract/2778476>)

CARE for reporting of clinical cases (<https://www.care-statement.org>)

AGREE for reporting clinical practice guidelines (<http://www.agreetrust.org/resource-centre/agree-reporting-checklist/>)

ARRIVE for reporting of animal pre-clinical studies (<https://arriveguidelines.org/arrive-guidelines>)

1. Word processors and format of manuscripts

A manuscript must be written in proper and clear English. Our preferred file format is DOCX or DOC. Manuscripts should be typed double-spaced on A4-sized paper, using 12 point font in English.

2. Abbreviation of terminology

Abbreviations should be avoided as much as possible. When they are used, full expression of the abbreviated words should be provided at the first use, with the abbreviation following in parentheses. Common abbreviations may be used, however, such as DNA. Abbreviations can be used if they are listed as a MeSH subject heading (<https://www.ncbi.nlm.nih.gov/mesh>).

3. Word spacing

1) Leave 1 space on each side when using arithmetic marks such as +, -, ×, etc.

Ex) 24 ± 2.5

Leave no space when using a hyphen between words.

Ex) intra-operative

2) When using parentheses, leave 1 space on each side.

3) When using brackets in parentheses, apply square brackets.

Ex) ([])

4. Citations

- 1) If a citation has 2 authors, write as “Hirota and Lambert”.
If there are more than 3 authors, apply “et al.” at the end of the first author’s surname.

Ex) Kim et al. [1]

- 2) Citations should be applied after the last word.

Ex) It is said that hypertension can be induced [1] and the way to injure the brain [2] is...

Ex) Choi and Kim [1] reported...

- 3) Apply citations before a comma or period.

Ex)is reported [1],

- 4) Several or coupled superscripts can be written as [1–5] or [1,3,5].

5. Arrangement of manuscript

The manuscript should be organized in the order of title, abstract, introduction, methods, results, discussion, acknowledgments, references, tables, figures, and figure legends. Figures should be uploaded as separate files. The title of each new section should begin on a new page. The conclusion should be included in the discussion section. Number pages consecutively, beginning with the first page of the manuscript. Page numbers should be placed in the middle of the bottom of the page. For survey-based clinical studies, the original survey document does not need to be included in the body of the manuscript but may be included as a supplement in an appendix.

6. Organization of manuscript

1) Original Article

- (1) Cover page (upload separately)

① Title

Title should be concise and precise. The first word should be capitalized. Drug names in the title should be written with generic names, not brand names. For the title, only the first letter of the first word should be capitalized.

Ex) Effect of smoking on bronchial mucus transport velocity under total intravenous anesthesia [○]

Ex) Effect of Smoking on Bronchial Mucus Transport Velocity under Total Intravenous Anesthesia ... [×]

Provide drug names as generic names, not product names.

Ex) In CPR, Isosorbide Dinitrate is, [○]

Ex) In CPR, Isosorbide Dinitrate (Isoket®) is, [×]

Ex) In CPR, Isoket® is, [×]

② Running title

A running title should be provided with no more than 40 characters, including letters and spaces in Korean, or 10 words in English. If this title is inappropriate, the Editorial Board may revise it.

③ Author information

First name, middle initial, and last name of each author, with their highest academic degree(s) (M.D., Ph.D., etc.), and institutional affiliations; make sure the names of and the order of authors as they appear on the Title Page and entered in the system match exactly.

④ Previous presentation at conferences

Title of the conference, date of presentation, and the location of the conference may be described.

(2) Manuscript

① Title and Running title (without author information)

It should be the same as the Cover page.

② Abstract

All manuscripts should contain a structured abstract that is written only in English. Authors should provide an abstract of no more than 250 words. It should contain 4 subsections: Background, Methods, Results, and Conclusions. Citation of references is not permitted in the abstract. A list of key words at least 6, with a maximum of 10 items, should be included at the end of the abstract. Key words should be selected from MeSH (<https://www.ncbi.nlm.nih.gov/mesh>), and these should be written in small letters with the first letter capitalized. Separate each word with a semicolon (;), and include a period (.) at the end of the last word.

Ex) Keywords: Carbon dioxide; Cerebral vessels; Oxygen; Spinal analgesia.

③ Introduction

The introduction should address the article’s purpose concisely and include background information relevant to the paper’s purpose.

④ Methods

The methods section should include sufficient details regarding the design, subjects, and methods of the research in order, as well as methods used for data analysis and control of bias in the study. Sufficient details must be provided in the methodology section of an experimental study so that others can further replicate it. The study design whether descriptive analysis, randomized controlled study, cohort study, or meta-analysis should be stated.

Materials and/or Participants: The materials used in the research should be clearly detailed to facilitate follow-up studies. Any materials purchased should be listed with the source or manufacturer. Research participants should also be precisely described with parameters such as age, sex, region, school, country, date of intervention period, occupation, etc. Reasons for inclusion or selection of participants should be explained. If a certain group was excluded, this should be explained as well. Questionnaires in non-English languages may also be included in the Appendix. Statistical analysis should be meticulously described. If reviewers want to analyze the data to confirm the results, the raw data may be provided to the editorial office. Computer programs used for the statistical analysis should be stated with the name, manufacturer, and software version used. Along with the statistical results, we encourage the inclusion of measurement error or uncertainty, such as listing confidence intervals in addition to providing P-values.

Institute and author names should be avoided.

When reporting experiments with human or animal subjects, the authors should indicate ethics statement whether they received approval from the Institutional Review Board for the study. If no IRB number is available, this should be discussed with the editor during the review process. When reporting experiments with animal subjects, the authors should indicate whether the Institutional Board supervised the handling of the animals for the Care and Use of Laboratory Animals. Demographic data should be included in the materials and methods section if applicable. As a rule, subsection titles are not recommended. If several study designs were used, then subtitles can be used without assigning numbers.

Ensure correct use of the terms sex (when reporting biological factors) and gender (identity, psychosocial or cultural factors), and, unless inappropriate, report the sex and/or gender of study participants, the sex of animals or cells, and describe the methods used to determine sex and gender. If the study was done involving an exclusive population, for example in only one sex, authors should justify why, except in obvious cases (e.g., prostate cancer).

Authors should define how they determined race or ethnicity and justify their relevance.

- **Units** Laboratory information should be reported using the International System of Units [SI], avail-

able at: <https://www.nist.gov/pml/special-publication-811>

< Exceptions >

- A. The unit for volume is “L,” while others should be written as “dl, ml, μ l”

Ex) 1 L, 5 ml

- B. The units for pressure are mmHg or cmH₂O. instead of Pascal.

- C. Use Celsius for temperature. oC

- D. Units for concentration are M, mM, μ M.

Ex) μ mol/L; [$\times$]

- E. When more than 2 items are presented, diagonal slashes are acceptable for simple units.

Negative exponents should not be used.

Ex) mg/kg/min [O], mg $\cdot$ kg⁻¹ $\cdot$ min⁻¹ [$\times$]

- F. Leave 1 space between number and units, except %, °C.

Ex) 5 mmHg

Ex) 5%, 36oC

- G. Units of time

Ex) hour: 1 h = 60 min = 3,600 s, day: 1 d = 24 h = 86,400 s

- **Machines and equipment**

According to the 11th edition of the American Medical Association, provide the model name and manufacturer’s name without the country.

For drug names, use generic names. If a brand name should be used, insert it in parentheses after the generic name. Provide® or TM as a superscript and the manufacturer’s name.

- **Ions**

Ex) Na⁺[O], Mg²⁺[O], Mg⁺⁺[$\times$], Mg⁺²[$\times$]

Ex) Premedicated magnesium [O]

Ex) Premedicated Mg²⁺ [O]

⑤ Results

Results should be presented in a logical sequence in the text, tables, and figures, giving the main or most important findings first. Do not repeat all the data provided in the tables or figures in the text; emphasize or summarize only the most important observations. Results can be sectioned by subsection titles but should not be numbered. The citation of tables and figures should be provided as Table 1 and Fig. 1.

Type or print each table on a separate page. Figures should be uploaded as separate tif, jpg, pdf, gif, ppt files.

⑥ Statistics

Precisely describe the methods of statistical analysis

and computer programs used. Mean and standard deviation should be described as mean $\pm$ SD, and mean and standard error should be written as mean $\pm$ SEM. Median and interquartile should be described as median (1Q, 3Q). When displaying P values, use a capital P and do not put a “-” between “P” and “value”.

- A. Describe the statistical tests employed in the study in enough detail so readers can reproduce the same results if the original data are available. The name and version of the statistical package should be provided.
 - B. Authors should describe the objective of the study and hypothesis appropriately. The primary/secondary endpoints are predetermined sensibly according to the objective of the study.
 - C. The characteristics of measured variables should determine the use of a parametric or nonparametric statistical method. When a parametric method is used, the authors should describe whether the basic statistical assumptions are met.
- For an analysis of a continuous variable, the normality of data should be examined. Describe the name and result of the particular method to test normality.
- D. When analyzing a categorical variable, an exact test or asymptotic method with appropriate adjustments should be used if the number of events and sample is small. The standard chi-squared test or difference-in-proportions test may be performed only when the sample size and the number of events are sufficiently large.
 - E. The *J Evid-Based Pract* strongly encourages authors to show confidence intervals, and it is not recommended to present the P value without showing the confidence interval. In addition, the uncertainty of estimated values, such as the confidence interval, should be described consistently in figures and tables.
 - F. Except for study designs that require a one-tailed test, for example, non-inferiority trials, the P values should be two-tailed. A P value should be expressed up to three decimal places (ex. $P = 0.160$ not as $P = 0.16$ or $P < 0.05$). If the value is less than 0.001, it should be described as “ $P < 0.001$ ” but never as “ $P = 0.000$.” For large P value greater than 0.1, the values can be rounded off to one decimal place, for example, $P = 0.1$, $P = 0.9$.
 - G. A priori sample size calculation should be described in detail. Sample size calculation must aim at preventing false negative results pertaining to the primary, instead of secondary, endpoint. Usually, the mean dif-

ference and standard deviation (SD) are typical parameters in estimating the effect size. The power must be equal to or greater than 80 percent. In the case of multiple comparisons, an adjusted level of significance is acceptable.

- H. When reporting a randomized clinical study, a CONSORT type flow diagram, as well as all the items in the CONSORT checklist, should be included. If limited in terms of the space of the manuscript, this information should be submitted as a separate file along with the manuscript.
- I. Results must be written in significant figures. The measured and derived numbers should be rounded off to reflect the original degree of precision. Calculated or estimated numbers (such as mean and SD) should be expressed in no more than one significant digit beyond the measured accuracy. Therefore, the mean (SD) of cardiac indices in patients measured on a scale that is accurate to 0.1 L/min/m² should be expressed as 2.42 (0.31) L/min/m².
- J. Except when otherwise stated herein, authors should conform to the most recent edition of the American Medical Association Manual of Style.

⑦ Discussion

The discussion should be described to emphasize the new and important aspects of the study, including the conclusions. Do not repeat in detail the results or other information that is provided in the introduction or the results section. Describe the conclusions according to the purpose of the study but avoid unqualified statements that are not adequately supported by the data. Conclusions may be stated briefly in the last paragraph of the discussion section.

⑧ ORCID (Open Researcher and Contributor ID)

All authors are required to provide a fully completed ORCID profile. ORCID registration is free and available to researchers worldwide through the ORCID website (<https://orcid.org>). Manuscripts submitted by authors who have not fully completed their ORCID profiles will not be considered for authorship and will be removed from the author list. Furthermore, if any listed author fails to meet this requirement, the manuscript will not proceed to the peer review process. An example ORCID profile is as follows: Owen Lee: <https://orcid.org/0000-0002-2117-1437>.

⑨ Authors' contributions

J Evid-Based Pract participates in the CRediT standard for author contributions. As such, the contribu-

tions of all authors must be described using the CRediT Taxonomy of author roles. For each of the categories below, please enter the initials of the authors who contributed in that category. If listing more than one author in a category, separate each set of initials with a space. If no author contributed to a category, you may leave that box blank.

The corresponding author is responsible for completing this information at submission, and it is expected that all authors will have reviewed, discussed, and agreed to their individual contributions before this time.

Examples of authors' contributions:

- Conceptualization: OL.
- Data curation: OL.
- Formal analysis: GJC.
- Funding acquisition: OL.
- Methodology: OL HK GJC.
- Project administration: GJC.
- Visualization: OL HK GJC.
- Writing – original draft: OL GJC.
- Writing – review & editing: OL HK GJC.

⑩ Conflict of Interest

Any conflicts should be disclosed here. This statement must be included regardless of the existence of conflicts of interest. If the authors have nothing to disclose, please state: "No potential conflict of interest relevant to this article was reported."

⑪ Funding

Financial support, including foundations, institutions, pharmaceutical and device manufacturers, private companies, intramural departmental sources, or any other support, should be described.

⑫ Data Availability Statement

J Evid-Based Pract has implemented a mandatory data sharing policy, requiring authors to submit raw data or data files at the time of manuscript submission for editorial review. Manuscripts submitted without the required dataset will not proceed to peer review. These data are essential for verifying the accuracy of the analysis and ensuring the reproducibility of results. Authors must upload data files in csv, xls, xlsx, or txt format. If an alternative file format is necessary, prior approval from the editorial office is required. If data sharing is restricted due to agreements with the data provider or other justified reasons, authors must consult with the editorial office before submission to discuss alternative data-sharing arrangements.

⑬ Acknowledgments

Persons or institutes that contributed to the manuscript but not sufficiently to be co-authors may be recognized.

⑭ Supplementary Materials

If supplementary materials are available, either to aid in reader understanding or because data are too abundant for inclusion in the main text, these may be included as supplementary data. Data files, as well as abstract recording, text, audio, or video files, can be added here.

⑮ References

- References should be obviously related to documents and should not exceed 50 in number. The number of references should not exceed 100 in reviews. However, the number of references has no limitation in systematic review and meta-analysis. References should be numbered consecutively in the order in which they are first mentioned in the text. Provide citations in the body text. All references should be listed in English, including author, title, name of journal, etc.
- The format for references follows the descriptions below. Otherwise, it follows the NLM Style Guide for Authors, Editors, and Publishers (Patrias, K. Citing medicine: the NLM style guide for authors, editors, and publishers [Internet]. 2nd ed. Wendling, DL, technical editor. Bethesda (MD): National Library of Medicine (US); 2007 [updated 2015 Oct 2; cited Year Month Day]. Available at: www.ncbi.nlm.nih.gov/books/NBK7256/).
- If necessary, the Editorial Board may request original documents for the references.
- The journal title should be listed according to the List of Journals Indexed for MEDLINE, available at: <https://www.nlm.nih.gov/archive/20130415/tsd/serials/lji.html>, or the List of KoreaMed Journals (journal browser of KoreaMed Services), available at: <http://koreamed.org/JournalBrowserNew.php>.
- Six authors can be listed. If there are more than 6 authors, only list 6 names with "et al."
- Provide the start and final page numbers of the cited reference.
- Abstracts of conferences may not be included in the references. The American Society of Anesthesiologists (ASA) refresher course lecture is not acceptable as a reference.
- Description format

A. Regular journal

- Author name. Title of article. Name of journal published year; volume: start page-final page.

Ex) Rosenfeld BA, Faraday N, Campbell D, Dorman T, Clarkson K, Siedler A, et al. Perioperative platelet activity of the effects of clonidine. *Anesthesiology* 1992; 79: 256-61.

Ex) Hirota K, Lambert DG. Ketamine: its mechanism(s) of action and unusual clinical uses. *Br J Anaesth* 1996; 77: 741-4.

Ex) Kang JG, Lee SM, Lim SW, Chung IS, Hahm TS, Kim JK, et al. Correlation of AEP, BIS, and OAA/S scores under stepwise sedation using propofol TCI in orthopedic patients undergoing total knee replacement arthroplasty under spinal anesthesia. *Korean J Anesthesiol* 2004; 46: 284-92.

- Journal article volume with supplement

Ex) Doherty JS, Froom SR, Gildersleve CD. Pediatric laryngoscopes and intubation aids old and new. *Pediatr Anaesth* 2009; 19 Suppl 1: 30-7.

- Journal article issue with supplement

Ex) Lee S, Han JW, Kim ES. Butyrylcholinesterase deficiency identified by preoperative patient interview. *Korean J Anesthesiol* 2013; 65(6 Suppl): S1-3.

B. Monographs

- Author. Book name. Edition. Place, press. Published year, pp (start page)-(End page).

- If reference page is only 1 page, mark 'p'.

- Note if it is beyond the 2nd edition.

Ex) Nuwer MR. Evoked potential monitoring in the operating room. 2nd ed. New York, Raven Press. 1986, pp 136- 71.

- Translated documents cannot be used as references. The original documents should be provided as references.

C. Chapter

Any separate author of a chapter should be provided.

Ex) Blitt C. Monitoring the anesthetized patient. In: *Clinical Anesthesia*. 3rd ed. Edited by Barash PG, Cullen BF, Stoelting RK: Philadelphia, Lippincott-Raven Publishers. 1997, pp 563-85.

D. Electronic documents

Ex) Grainge MJ, Seth R, Guo L, Neal KR, Coupland C, Vryenhoef P, et al. Cervical human papillomavirus screening among older women. *Emerg Infect Dis* [serial on the Internet]. 2005 Nov [2005 Nov 25]. Available from wwwnc.cdc.gov/eid/article/11/11/05-0575_article.

E. Online journal article

Ex) Sampson AL, Singer RF, Walters GD. Uric acid lowering therapies for preventing or delaying the progression of chronic kidney disease. *Cochrane Database Syst Rev* 2017; 10: CD009460.

F. Advance access article

Ex) Baumbach P, Gotz T, Gunther A, Weiss T, Meissner W. Chronic intensive care-related pain: Exploratory analysis on predictors and influence on health-related quality of life. *Eur J Pain* 2017. Advance Access published on Nov 5, 2017. doi:10.1002/ejp.1129.

The reference style for *J Evid-Based Pract* is conveniently available as an out-of-the-box style within both EndNote and RefWorks.

⑩ Tables

Only one table is to be drawn per page in the order cited in the text.

The title of the table is to be in English and written at the top of the table in the form of a phrase.

Words in the table excluding the title should use capital letters for the first word, and the following words are to be written in small letters.

For demographic data, gender is recorded as M/F, age as yr, (if necessary, use days or months in children) without decimal point. The “±” sign within the table is to be aligned with the rows above and below.

Footnotes are to be written in the following order: “Values are mean ± SD (or SEM) or median (1Q, 3Q)”, the explanations for the groups and the abbreviations in order of appearance, and statistics. Abbreviations apart from internationally recognized abbreviations are to be explained with their full spellings at the bottom of the table. Full spellings are to be presented even for repeated abbreviations for each table in order of appearance.

Significance marks are to conform to the Vancouver style (Uniform Requirements for Manuscripts Submitted to Biomedical Journals. *JAMA* 1997; 227: 927-34). In other words, these must be in the order of *, †, ‡, §, //, ¶, **, ††, ‡‡ and written as superscripts.

⑪ Legends for figures and photographs

All of the figures and photographs should be described in the text separately.

The description order is the same as in the footnotes in tables and should be in recognizable sentences.

Define all abbreviations every time they are repeated.

(3) Figures and Photographs

- ① JBEP encourages authors to use color to increase the clarity of figures. Please note that color figures are used without charge for online reading. However, since it will be charged upon the publication, authors may choose to use colors only for online reading.
- ② Standard colors should be used (black, red, green, blue, cyan, magenta, orange, and gray). Avoid colors that are difficult to see on the printed page (e.g., yellow) or are visually distracting (e.g., pink). Figure backgrounds and plot areas should be white, not gray. Axis lines and ticks should be black and thick enough to frame the image clearly. Axis labels should be large enough to be easily readable, and printed in black.
- ③ Figures should be uploaded as separate tif, jpg, pdf, gif, or ppt files. The width of figure should be 84 mm (one column). The contrast of photos or graphs should be at least 600 dpi. The contrast of line drawings should be at least 1,200 dpi. Number figures as "Fig. (Arabic numeral)" in the order of their citation (ex. Fig. 1).
- ④ Photographs should be submitted individually. If Fig. 1 is divided into A, B, C, and D, do not combine it into 1, but submit each of them separately. Authors should submit line drawings in black and white.
- ⑤ In horizontal and vertical legends, the letter of the first English word should be capitalized.
- ⑥ Connections between numbers should be denoted by "-", not "~". Do not space the numbers (ex. 2-4).
- ⑦ An individual should not be recognizable in photographs or X-ray films unless written consent has been obtained from the subject and is provided at the time of submission.
- ⑧ Pathological samples should be pictured with a measuring stick.

2) Review

This review article synthesizes previously published material into an integrated presentation of our current understanding of a topic. Review articles should describe aspects of a topic in which scientific consensus exists, as well as aspects that remain controversial and are the subject of ongoing scientific disagreement and research. Review articles are invited only by editorial board. If authors want to submit an unsolicited review article, please contact editorial office (ksebm.office@gmail.com). Review articles should include unstructured abstracts written in English equal to or less than 250 words. The organization should be in order of abstract, introduction, text following each title, conclusion and references.

Figures and tables should be provided in English. Body text should not exceed 30 A4-sized pages, and the number of figures and tables should each be less than 6. However, if necessary, the number of pages, the number of figures and tables can be added in accordance with the decision of the editorial committee.

3) Systematic review and meta-analysis

Systematic review and meta-analysis are considered as an original article. Systematic reviews are systematic, critical assessments of literature and data sources in order to answer a specific question, and/or includes a statistical technique leading to a quantitative summary of results and examining sources of differences in results among studies, if any. The subtitle should include the phrase "A systematic review" and/or "A Meta-analysis." Organization of systematic review and meta-analysis: Same as original article, except,

- All systematic reviews and meta-analyses should be registered at an appropriate online public registry (eg, PROSPERO; <http://www.crd.york.ac.uk/PROSPERO/>), and registration information should be included with the submission. Authors of reports of meta-analyses of clinical trials should submit the PRISMA flow diagram. The PRISMA checklist should be submitted as a separate file along with the manuscript. For information regarding PRISMA guidelines, please visit <http://www.prisma-statement.org> or EQUATOR Network (<https://www.equator-network.org/home/>). Systematic reviews and meta-analyses of observational studies in epidemiology should be reported according to MOOSE guidelines. For more information regarding MOOSE guidelines, please visit <http://www.equator-network.org/reporting-guidelines/meta-analysis-of-observational-studies-in-epidemiology-a-proposal-for-reporting-meta-analysis-of-observational-studies-in-epidemiology-moose-group/>.
- Number of references has no limitation in systematic review and meta-analysis.

4) Case Report

A case report is almost never a suitable means to describe the efficacy of a treatment or a drug; instead, an adequately powered and well-controlled clinical trial should be performed to demonstrate such efficacy. The only context in which a case report can be used to describe efficacy is in a clinical scenario, or population, that is so unusual that a clinical trial is not feasible. Case reports of humans must state in the text that informed consent to publication was obtained from the patient or guardian. Copies of written informed consents should be kept. If necessary, the editor or reviewers

may request copies of these documents. If these steps are impossible, Institutional Review Board approval should be obtained prior to submission. The rarity of a disease condition is itself not an acceptable justification for a case report. Statement describing compliance with CARE for reporting of clinical cases (<https://www.care-statement.org>) guideline is recommended.

- (1) Cover page: Same as that for clinical and experimental studies.
- (2) Abstract: All case reports should contain a structured abstract that is written only in English. Provide an abstract of no more than 150 words. It should contain 3 subsections: Background, Case, and Conclusions. A list of keywords, with a Minimum of 6, should be included at the end of the abstract.
- (3) Introduction: Should not be separately divided. Briefly describe the case and background without a title.
- (4) Case report: Describe only the clinical information that is directly related to the diagnosis and anesthetic management.
- (5) Discussion: Briefly discuss the case, and state conclusions at the end of the case. Do not structure the conclusion section separately.
- (6) References: The number of references should be less than 20. However, if necessary, the number of reference

can be added in accordance with the decision of the editorial committee.

- (7) Tables and figures: Proportional to those for clinical and experimental studies.

5) Letter to the Editor

Letter to the Editor should include brief constructive comments that concern previously published articles and interesting cases. Letters to the Editor should be submitted no more than 3 months after the paper has been published.

- (1) Cover pages should be formatted in the same way as those of clinical research papers. The corresponding author should be the first author. A maximum of five authors is allowable.
- (2) The body text should not exceed 1,000 words and should have no more than 5 references. A figure or a table may be used.
- (3) Letters may be edited by the Editorial Board, and if necessary, responses by the author of the subject paper may be provided.

6) Editorial

Editorial is invited by the editorial committee and should be commentaries on articles recently published in the *J Evid-Based Pract*, and can be described in free style.

Author's checklist

- ☐ Manuscript in MS-WORD (DOC, DOCX) format.
- ☐ Double-spaced typing with 12-point font.
- ☐ Sequence of title page, abstract and keywords, introduction, methods, results, discussion, and conclusions, acknowledgments, references, tables, and figure legends. All pages and manuscript text with line should be numbered sequentially, starting from the abstract.
- ☐ Title page with article title, authors' full name(s) and affiliation(s), address for correspondence (including telephone number, e-mail address, and fax number), running title (less than 50 characters), and acknowledgments, if any.
- ☐ Abstract in structured format up to 300 words for original articles. Keywords (up to 5) from the MeSH list of Index Medicus.
- ☐ All table and figure numbers are found in the text.
- ☐ Figures as separate files, in TIFF, JPG, GIF, or PPT format.
- ☐ References listed in proper format. All references listed in the reference section are cited in the text and vice versa.

Copyright transfer statement

☐ Original Article ☐ Review ☐ Systematic Review ☐ Clinical Trial ☐ Clinical Practice Guideline
☐ Case Report ☐ Letter to the Editor ☐ Editorial ☐ Others

Article title:

Author(s):

(In identical order to the electronic submission and the corresponding author should be underlined)

Journal: Journal of Evidence-Based Practice

In the event that the above manuscript is accepted for publication in *Journal of Evidence-Based Practice*, the copyright is transferred to the Korean Society of Evidence-Based Medicine.

The author shall have the right to use all or part of the Work to revise, adapt, prepare derivative works, present orally, or distribute, provided that such use is for the personal, non-commercial benefit of the author. With written consent from *Journal of Evidence-Based Practice*, the author may use content from the Work in other works, provided that a full acknowledgment is made to the original source of the material, including the journal name, volume, issue, page numbers, year of publication, and title of the article.

All co-authors of the Work certify that they have participated sufficiently in the intellectual content, the analysis of data (if applicable), and the writing of the Work to take public responsibility for it. Each author of the Work certifies that none of the material in the Work has been previously published, included in another work, or is currently under consideration for publication elsewhere. Each author certifies that this Work has not been accepted for publication elsewhere, nor has assigned any right or interest in the Work to any third party.

All funding sources supporting the Work and all institutional or corporate affiliations of the authors and commercial associations (e.g., consultancies, stock ownership, equity interests, patent-licensing arrangements, etc.) that might pose a conflict of interest in connection with the Work should be acknowledged in a footnote on the front cover of the Work.

I/we give consent to the above statements. (It is the responsibility of the Corresponding Author to collect the signatures of all authors before sending the form to the Editorial Office.)

Author's Signature	Date	Author's Signature	Date
Author's Signature	Date	Author's Signature	Date
Author's Signature	Date	Author's Signature	Date