

Evaluation of a Problem-Based Learning Program's Effect on Artificial Intelligence Ethics among Japanese Medical Students: Mixed-Methods Study

Yuma Ota, Yoshikazu Asada, Saori Kubo, Takeshi Kanno, Machiko Saeki yagi,
Yasushi Matsuyama

Submitted to: JMIR Medical Education
on: September 22, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 5

Supplementary Files..... 33

 Figures 34

 Figure 1..... 35

 Figure 2..... 36

 Figure 3..... 37

 Figure 4..... 38

Evaluation of a Problem-Based Learning Program's Effect on Artificial Intelligence Ethics among Japanese Medical Students: Mixed-Methods Study

Yuma Ota¹ MSN; Yoshikazu Asada² PhD; Saori Kubo³ PhD; Takeshi Kanno^{4,5} PhD; Machiko Saeki yagi² PhD; Yasushi Matsuyama² PhD

¹ Medical Education Center Jichi Medical University Graduate School Shimotsuke-shi, Tochigi JP

² Medical Education Center Jichi Medical University Tochigi JP

³ Graduate School of Education Tohoku University Sendai JP

⁴ R & D Division of Career Education for Medical Professionals Medical Education Center Jichi Medical University Tochigi JP

⁵ Division of Gastroenterology Graduate School of Medicine Tohoku University Sendai JP

Corresponding Author:

Yasushi Matsuyama PhD

Medical Education Center
Jichi Medical University
3311-1 Yakushiji, Shimotsuke-shi
Tochigi
JP

Abstract

Background: The impact of the rapid advancement of artificial intelligence (AI) on the medical domain is significant. Accordingly, it is necessary to integrate AI education into medical school curricula. However, the integration of AI education remains limited. A key challenge to this integration is the discrepancy between medical students' positive perceptions of AI and their actual competencies, with research in Japan identifying specific gaps in the students' competencies in understanding regulations and discussing ethical issues.

Objective: This study evaluates the effectiveness of an educational program designed to improve medical students' competencies in understanding legal and ethical AI-related issues. It addresses the following research questions: (1) Does this educational program improve students' knowledge of AI and its legal and ethical issues, and what is each program element's contribution to this knowledge? (2) How does this educational program qualitatively change medical students' thoughts on these issues from an abstract understanding to a concrete and structured thought process?

Methods: This study adopted a mixed-methods, convergent parallel design using a single-group pretest/posttest framework involving 118 fourth-year medical students. The one-day intervention comprised a lecture and Problem-Based Learning (PBL) session centered on a clinical case. Further, a 24-item multiple-choice questionnaire (MCQ) was administered at three time points (pretest, mid-test, and posttest), and descriptive essays were collected before and after the intervention. Data were analyzed using linear mixed-effects models; the Wilcoxon signed-rank test; and text mining, including cooccurrence network analysis with Jaccard coefficients. An optional survey on student perceptions (ARCS model) was conducted (N=76, response rate = 64.4%), as well.

Results: The program caused a significant increase in objective knowledge from the pretest (median 17(15-18)) to posttest (median 19(17-21); $\chi^2 = 1.42$, $P < 0.001$) periods. However, no statistically significant difference was noted between the score gains of the lecture phase and PBL phase ($P = .54$). Further, qualitative analysis revealed the significant transformation of cooccurrence network structures (Jaccard coefficients: 0.116 and 0.121) from fragmented clusters to integrated networks. Word frequency analysis revealed the increased use of professional and ethical terms (e.g., bias and personal information). The optional survey indicated that students' confidence (mean = 3.78, standard deviation (SD) = 0.87) was significantly lower than their perception of the program's relevance (mean = 4.20, SD = 0.71; $P < 0.001$).

Conclusions: The PBL program effectively improved overall knowledge and, more importantly, qualitatively transformed students' thinking about AI ethics from an abstract level to a concrete and structured thought process based on clinical context.

The discrepancy between the quantitative and qualitative results of the study suggests MCQs' insufficiency in assessing the higher-order skills fostered by PBL. Finally, this mixed-methods study indicates that PBL is an effective pedagogical method for AI ethics education.

(JMIR Preprints 22/09/2025:84535)

DOI: <https://doi.org/10.2196/preprints.84535>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <https://preprints.jmir.org/preprint/84535>

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <https://preprints.jmir.org/preprint/84535>

Original Manuscript

Evaluation of a Problem-Based Learning Program's Effect on Artificial Intelligence Ethics among
Japanese Medical Students: Mixed-Methods Study



Abstract

Background: The impact of the rapid advancement of artificial intelligence (AI) on the medical domain is significant. Accordingly, it is necessary to integrate AI education into medical school curricula. However, the integration of AI education remains limited. A key challenge to this integration is the discrepancy between medical students' positive perceptions of AI and their actual competencies, with research in Japan identifying specific gaps in the students' competencies in understanding regulations and discussing ethical issues.

Objective: This study evaluates the effectiveness of an educational program designed to improve medical students' competencies in understanding legal and ethical AI-related issues. It addresses the following research questions: (1) Does this educational program improve students' knowledge of AI and its legal and ethical issues, and what is each program element's contribution to this knowledge? (2) How does this educational program qualitatively change medical students' thoughts on these issues from an abstract understanding to a concrete and structured thought process?

Methods: This study adopted a mixed-methods, convergent parallel design using a single-group pretest/posttest framework involving 118 fourth-year medical students. The one-day intervention comprised a lecture and Problem-Based Learning (PBL) session centered on a clinical case. Further, a 24-item multiple-choice questionnaire (MCQ) was administered at three time points (pretest, mid-test, and posttest), and descriptive essays were collected before and after the intervention. Data were analyzed using linear mixed-effects models; the Wilcoxon signed-rank test; and text mining, including cooccurrence network analysis with Jaccard coefficients. An optional survey on student perceptions (ARCS model) was conducted (N=76, response rate = 64.4%), as well.

Results: The program caused a significant increase in objective knowledge from the pretest (median 17(15-18)) to posttest (median 19(17-21); $\beta = 1.42$, $P < 0.001$) periods. However, no statistically significant difference was noted between the score gains of the lecture phase and PBL phase ($P = .54$). Further, qualitative analysis revealed the significant transformation of cooccurrence network

structures (Jaccard coefficients: 0.116 and 0.121) from fragmented clusters to integrated networks. Word frequency analysis revealed the increased use of professional and ethical terms (e.g., bias and personal information). The optional survey indicated that students' confidence (mean = 3.78, standard deviation (SD) = 0.87) was significantly lower than their perception of the program's relevance (mean = 4.20, SD = 0.71; $P < 0.001$).

Conclusions: The PBL program effectively improved overall knowledge and, more importantly, qualitatively transformed students' thinking about AI ethics from an abstract level to a concrete and structured thought process based on clinical context. The discrepancy between the quantitative and qualitative results of the study suggests MCQs' insufficiency in assessing the higher-order skills fostered by PBL. Finally, this mixed-methods study indicates that PBL is an effective pedagogical method for AI ethics education.

Keywords: Education, Medical, Undergraduate; Artificial Intelligence; Ethics, Medical; Problem-Based Learning; Mixed-Methods Research; Text Mining

Introduction

The rapid advancement of artificial intelligence (AI) has significantly affected the medical domain, which utilizes applications ranging from diagnostic support to treatment selection [1-3]. This necessitates the integration of AI education into medical school curricula to ensure future physicians have the necessary competencies to use these technologies effectively and ethically [4-6]. Both educators and students recognize this necessity, and many studies support the implementation of systematic AI education [7,8]. However, despite the establishment of international recommendations, the integration of AI education into medical school curricula remains limited [3,6]. One of the key challenges in AI education identified by earlier research is the discrepancy between medical students' perceptions of and competencies in AI [6,7,9]. Recent systematic reviews consistently indicate that despite generally maintaining a positive attitude toward AI's healthcare potential, students have a moderate to low level of objective knowledge and practical skills [7]. This gap between the positive perceptions and insufficient competency acquisition of students indicates the need to integrate systematic and effective educational programs into the medical curriculum [9,10,11].

Furthermore, AI education poses not only technical challenges but also legal and ethical ones [1,2,12]. The use of AI can give rise to new legal and ethical issues, such as the risk of algorithmic bias resulting in patient discrimination, challenges related to patient privacy, and problems associated with accountability for AI-driven decisions [2,3,13]. Further, expert consensus studies indicate that nontechnical competencies such as ethics, law, communication, and collaboration are considered more important than advanced technical skills such as programming for physicians [8,11]. However, practical, case-based ethics education and standardized frameworks to teach these important topics are currently lacking [3,12,13].

This global context is highly relevant in Japan. In response to these trends, Japan's Ministry of Education, Culture, Sports, Science and Technology (MEXT) revised the national Model Core

Curriculum for medical education in 2022, establishing "Competency in Utilizing Information Science and Technology" as a core competency for all future physicians [14,15]. This necessitated an investigation of the specific educational needs of Japanese medical students. A recent national cross-sectional study of sixth-year medical students in Japan clarified the acquisition status of specific competencies [16]. The study revealed particularly low self-assessment in three key learning objectives: (1) understanding the regulations, laws, and guidelines for information science and technology (IST) adoption in medicine; (2) discussing ethical issues, such as social disparities; and (3) understanding and discussing the potential applications of specific technologies, including AI, in medical care [16]. These results highlight the necessity of implementing focused educational interventions to address these gaps among Japanese medical students. To address these identified gaps, we chose a Problem-Based Learning (PBL) framework for our educational intervention. This decision was directly informed by the findings of the aforementioned national study [16], which not only highlighted these specific areas of low confidence but also revealed a significant positive association between prior PBL experience and students' self-assessed competence in these very domains. This suggests that the PBL methodology—which promotes active, case-based, and collaborative problem-solving—is a particularly effective approach for navigating the complex, non-technical challenges of AI ethics and law.

Accordingly, this study aims to evaluate the effectiveness of an educational program designed to improve medical students' competencies in addressing the legal and ethical issues associated with AI use. To this end, we developed and implemented a PBL program centered on a clinical case requiring fourth-year medical students to navigate AI-related ethical dilemmas. The use of the term "artificial intelligence" by Japan's Model Core Curriculum highlights AI's development period [14,15]. However, the educational program developed in this study considers recent technological advancements and the societal landscape to primarily address the ethical challenges posed by Generative AI. Therefore, although this study uses the term AI in a broader sense in alignment with

earlier research and the core curriculum, its content and examples mainly focus on Generative AI. Finally, we address the following research questions:

Research Question 1 (RQ1): Does this educational program improve students' knowledge of AI and its legal and ethical issues? Furthermore, which program elements improve knowledge, and how?

Research Question 2 (RQ2): How does the educational program qualitatively change medical students' thoughts on the legal and ethical issues of AI? In particular, does their thinking shift from an abstract understanding to a more concrete and structured thought process that is based on the clinical context?

Methods

Study Design and Participants

This study adopted a mixed-methods, convergent parallel design using a single-group pretest/posttest framework to evaluate an educational intervention. The participants were 124 fourth-year medical students at Jichi Medical University who undertook the intervention program as a mandatory part of their curriculum during the 2025 academic year. The program was conducted immediately before the students commenced their full-time clinical clerkship; this time was selected because they were expected to soon encounter ethical and legal issues related to AI's adoption in clinical practice.

Educational Intervention

The intervention in this study was a 1-day (approx. 7-hour) program based on the PBL model [17]. The program's learning objectives were selected such that they address the three key competency gaps (understanding regulations, discussing ethical issues, and understanding AI applications) that had low self-assessment scores among Japanese medical students in an earlier national cross-sectional study. The program required students to achieve the following three competencies based on a rubric proposed by earlier research [18]:

1. For standard clinical settings and cases, be able to take action based on regulations (laws,

guidelines, etc.) associated with IST's use in medicine at a certain rate.

2. Be able to exemplify and discuss the ethical issues, such as the social disparities caused by the digital divide, potentially arising from the use of digital information and scientific technology in medicine and welfare.

3. Be able to understand the IST (AI) related to medicine and discuss it with the support of practitioners.

PBL Case and Program Structure

At the beginning of the PBL-based program, a video depicting a challenging clinical scenario was shown: A patient who is diagnosed with "muscle pain" for his left shoulder pain by a medical AI consultation app is convinced of the diagnosis. However, his anxious son is concerned about the app's reliability (since it lists acute myocardial infarction as the second possibility) and asks a medical student on clinical rotation to explain the AI diagnosis' credibility. The case was designed to encompass multiple key AI ethics themes, including physician accountability, potential AI bias and safety issues, and empathetic communication.

The program was conducted over one complete day with the following structure (Table 1): (1) At the beginning of the program, the learning objectives and the case video were presented. The case involved a scenario in which a medical student on clinical rotation was asked by a patient's anxious son to explain the reliability of the medical AI consultation app that had been used to diagnose the patient's left shoulder pain as muscle pain, a diagnosis the patient firmly believed. This case encompassed multiple key themes in AI ethics, including physician accountability, potential AI bias and safety issues, and empathetic communication. Subsequently, (2) a preprogram assessment (pretest: a multiple-choice questionnaire (MCQ) and descriptive questions) was conducted. Then, (3) the foundational knowledge of AI was detailed in Lecture 1, which referenced internationally acknowledged AI education content for medical students [3]. This lecture supported the acquisition of sufficient foundational knowledge to address the PBL case. It provided an overview of the

fundamentals of AI, including its definition, strengths, machine and deep learning mechanisms, and the benefits and challenges of AI's introduction into healthcare. Immediately after Lecture 1, (4) the second assessment (mid-test: MCQ alone) was administered.

The PBL process proceeded in the following stages: (5) PBL learning process confirmation, (6) problem clarification, (7) self-directed learning (including a lunch break), and (8) a 70-minute problem-solving groupwork session. PBL was conducted in 16 small groups of 7–8 students each. The session comprised problem clarification, self-directed learning, and problem-solving phases and included a final 70-minute groupwork session where students used shared Google slides to collaboratively devise an explanation for the patient and his son. Five tutors facilitated the process using a standardized tutor guide and, to enhance discussion quality, key domestic AI guidelines [19, 20] were distributed as groupwork materials.

At the end of the program, (9) in Lecture 2, an instructor reviewed the case to consolidate important concepts regarding AI's legal and ethical issues. Immediately thereafter, (10) the final assessment (posttest: an MCQ and descriptive questions) was administered. A detailed description of the PBL program's development and theoretical underpinnings has been published [21].

Table1. Overview of the program components and schedule

Content	Time	Activity format
(1) Introduction: Presentation of Learning Objectives and Case	15 minutes	Individual
(2) Assessment 1 (Pre-test): MCQ and Descriptive Questions	35 minutes	Individual
(3) Lecture 1: Foundational Knowledge of AI	35 minutes	Individual
(4) Assessment 2 (Mid-test): MCQ	15 minutes	Individual
(5) PBL Session 1: Confirmation of Schedule and Learning Process	5 minutes	Group
(6) PBL Session 2: Problem Clarification	20 minutes	Group
(7) PBL Session 3: Self-Learning (including lunch break)	120 minutes	Individual

(8) PBL Session: Problem-Solving Group Work	70 minutes	Group
(9) Lecture 2: Case Review and Debriefing	40 minutes	Individual
(10) Assessment 3 (Post-test): MCQ and Descriptive Questions	35 minutes	Individual

Data Collection

To **evaluate** the program's effectiveness, we collected both quantitative and qualitative data using the university's learning management system.

Using the MCQ

A 24-item MCQ, with its content aligned with Lecture 1 materials, was used to assess students' knowledge. Each item scored 1 point for a correct answer, the total possible score being 24 points. The questionnaire comprised eight subscales, each with three items, covering the mechanisms of AI, the strengths of AI, machine learning, deep learning, overfitting, prompts, personal information protection, and challenges associated with AI utilization. To assess knowledge acquisition over time, the same set of questions was administered at three time points: before the intervention (pretest), after the lecture (mid-test), and after the PBL program's completion (posttest). To minimize recall bias, the order of the questions and their options was randomized at each administration.

Descriptive Essay Questions

Two descriptive essay questions were administered before the intervention (pretest) and after the program (posttest): "A patient literally believes the results diagnosed by a personal AI consultation tool. As a medical student, how will you explain the dangers of believing the results diagnosed by AI?" (Question 1) and "When using AI as a physician, what points do you think require caution?" (Question 2). These questions were designed to evaluate changes in students' patient communication strategies and their understanding of the professional and ethical responsibilities

associated with AI use.

Program Perception Survey

After the program, an optional survey was administered to assess students' perceptions of the program based on the four ARCS model components: attention, relevance, confidence, and satisfaction [22]. Students rated their agreement with statements for each component on a 5-point Likert scale (ranging from 1=strongly disagree to 5=strongly agree).

Data Analysis

All statistical analyses were performed using R (version 4.4.3)[23].

MCQ Data

The normality of data distribution for score gains was assessed using the Shapiro–Wilk test to determine the appropriate statistical test to compare educational phases. To assess the overall trajectory of the increase in score across the time points pretest, mid-test, and posttest, a linear mixed-effects model was used. To compare the score gains between the lecture phase (mid-test–pretest) and the PBL phase (posttest–mid-test), a paired-samples t-test was performed. Any violation of the assumption of normality was prespecified to result in the use of the nonparametric Wilcoxon signed-rank test as an alternative.

Descriptive Essay Data

To quantitatively and objectively analyze the content of free-text data, a text-mining approach utilizing several R packages was adopted. The analysis targeted the free-text responses to two questions at two time points: Question 1 (patient explanation) at pre- and posttest and Question 2 (physician precautions) at pre- and posttest. Further, data handling, aggregation, and visualization were performed using the dplyr, tidyr, tidytext, widyr, igraph, and ggraph R packages [24]. The analytical procedure was as follows:

1. Morphological analysis for word tokenization and normalization: To segment Japanese free-text into analyzable word units, morphological analysis was performed using the RMeCab

package (version 1.1.4) [25], which interfaces with the open-source engine MeCab, utilizing the standard IPA dictionary (ipadic). The analysis was limited to content words (nouns, verbs, and adjectives), with verbs and adjectives being normalized to nouns. Nouns alone are typically used as morphemes in network analysis [26].

2. Word frequency analysis: Extracted and normalized words were used to calculate the word frequencies for the responses to each question at each time point (Question 1 at pretest, Question 1 at posttest, Question 2 at pretest, and Question 2 at posttest) to identify the most frequently used terms.

3. Comparative word frequency analysis: To capture changes in descriptive content, word frequencies were compared among time points for each question (pretest vs. posttest for Question 1, and pretest vs. posttest for Question 2). By calculating the difference between posttest and pretest frequencies, words with increased or decreased frequencies were extracted.

4. Cooccurrence network analysis: To structurally clarify the contextual relationships among words irrespective of their frequencies, a cooccurrence network analysis was performed. In this study, cooccurrence involves capturing the semantic proximity of a pair of words appearing adjacently within the same sentence. Once text was segmented into sentence units, adjacent word pairs were extracted and their frequencies (the number of students who used each word pair) were calculated using the widyr package. Subsequently, a network diagram was created using the words mentioned by five or more students as nodes and their cooccurrence relationships as edges, with edge thickness indicating the relationship's strength.

The Jaccard coefficient was used as a similarity metric to quantify the structural changes in these networks. The Jaccard coefficient measures the similarity between two sets by taking their intersection and dividing it by their union [27]. In this study, the coefficient was applied to compare entire network structures and was defined as the number of shared cooccurrence pairs (edges) divided by the total number of unique pairs across both pre- and posttest networks. We applied this metric because it assesses the number of common cooccurrence pairs without being affected by each

pair's frequency. Accordingly, we could determine the extent to which the network structure changed over time, which indicated a shift in conceptual relationships, rather than a simple change in discourse volume. We identified the numbers of common pairs (present at both time points) and unique pairs (present at only one time point), as well.

Program Perception Survey Data

For the optional ARCS survey data, descriptive statistics (mean and standard deviation) were calculated for each of the four components. To clarify the differences between components, repeated-measures analysis of variance (ANOVA) was conducted. Further, pairwise t-tests with a Bonferroni correction were used for post-hoc analysis.

Ethical Considerations

The study's description stated that participation was voluntary and would not affect students' grades. Only those medical students who agreed to participate accessed the questionnaire. The study was approved by the Ethics Review Committee of Jichi Medical University School of Medicine, Shimotsuke, Japan (approval number: 24-137).

Results

Participant Characteristics

In this study, 118 fourth-year medical students participated. All participants provided complete data for both the MCQ and descriptive essay questions, which were the main aspects analyzed in this study. The participants were 71 male (60.2 %) and 47 female (39.8 %) students, all being below 30 years of age.

MCQ Scores: Descriptive Statistics and Overall Effectiveness

The MCQ's internal consistency and reliability were assessed using the Cronbach's alpha value at each time point. The alpha coefficients for the total score were 0.651 at pretest, 0.765 at mid-test, and 0.782 at posttest. The alpha coefficients for the eight subscales ranged from 0.631 to

0.760, indicating acceptable-to-good reliability. Table 2 depicts the detailed Cronbach's alpha values for the total score and all subscales.

Table2.Cronbach's Alpha Coefficients for the MCQ at Each Time Point

Factor	Pretest	Mid-test	Posttest
Strengths of AI	0.631	0.662	0.692
AI Architecture	0.639	0.744	0.724
AI output issues	0.743	0.732	0.754
Prompts	0.657	0.707	0.696
Personal Data Leakage	0.76	0.708	0.77
Machine Learning	0.666	0.751	0.726
Deep Learning	0.704	0.729	0.757
Overfitting	0.716	0.733	0.743
Total Score	0.651	0.765	0.782

As per the Shapiro–Wilk test, the MCQ scores at none of the three time points were normally distributed ($P < .001$ for all values). Therefore, descriptive statistics were presented as median values (Q1-Q3). The median total MCQ score increased across the time points, from 17 (15-18) at pretest to 18 (16-20) at mid-test and 19 (17-21) at posttest (Table 3).

Table3. Median Scores and Quartiles (Q1, Q3) of the MCQ at Each Time Point

Factor	pretest	mid-test	posttest
Strengths of AI	2 (2-3)	2 (2-3)	2 (2-3)
Mechanisms of AI	2 (2-3)	2 (2-3)	3 (2-3)
Challenges Associated with AI Utilization	2 (2-3)	2 (2-3)	2 (2-3)
Prompts	2 (1-2)	2 (1-3)	2 (1-3)
Personal Information Protection	3 (3-3)	3 (3-3)	3 (3-3)
Machine Learning	2 (1-2)	2 (1-3)	2 (1-3)
Deep Learning	1 (1-2)	2 (1-3)	2 (1-3)
Overfitting	2 (1-3)	2 (1-3)	2 (1-3)
Total Score	17 (15-18)	18 (16-20)	19 (17-21)

A linear mixed-effects model was used to assess the trajectory of this increase in scores.

Results (Table 4) highlight the stepwise effect of the program. For instance, a statistically significant increase of 0.75 points from the baseline score at pretest was noted after the lecture phase at mid-test ($\beta = 0.75$, $P = .01$). Furthermore, the end of program (posttest) score significantly exceeded the baseline score by 1.42 points ($\beta = 1.42$, $P < 0.001$).

Table4. Fixed Effects from Linear Mixed-Effects Model Predicting Total Score

Effect	Estimate (β)	SE	df	t	p
Intercept (pretest)	16.86	0.34	179	49.55	< .001
T2 (mid-test)	0.75	0.27	234	2.79	0.006
T3 (posttest)	1.42	0.27	234	5.30	< .001

Model: Score $\sim$ Time + (1 | ID); N = 354 observations from 118 students.

To address the quantitative research question (RQ1) regarding the PBL phase's additional effect, the Wilcoxon signed-rank test was performed to compare score gains. The test found no statistically significant difference between the score gain during the lecture phase (mid-test–Pre-test) and that during the PBL phase (Post-test–mid-test) for either the total score ($V = 3084.5$, $P = .54$) or any of the eight subscales (Table 5).

Table5. Wilcoxon Signed-Rank Test Comparing Score Gains Between Lecture (Mid-test - Pre-test) and PBL (Post-test - Mid-test) Phases

Factor	V Statistic	p-value
Strengths of AI	203.0	0.058
Mechanisms of AI	1733.5	0.214
Challenges Associated with AI Utilization	1347.0	0.820
Prompts	1601.5	0.759
Personal Information Protection	1044.0	0.545
Machine Learning	981.0	0.612
Deep Learning	120.0	0.311
Overfitting	782.0	0.895
Total Score	3084.5	0.541

Qualitative Research Question

To address RQ2, free-text responses to Question 1 (patient explanation) and Question 2 (physician precautions) were analyzed by comparing word frequencies and cooccurrence network structures between pre- and posttests.

Responses to Question 1

After the intervention, a significant increase in the number of students using words related to AI mechanisms, limitations, and applications was noted. In particular, a high number of students used "Learning" (increase of 18 students), "Input" (+17), "Bias" (+15), and "Utilization" (+13). Contrastingly, the use of words describing the clinical case's initial facts was low. Only a few students used words such as "Symptoms" (decrease of 11 students), "Examination" (-11), and "Explanation" (-9). Table 6 depicts the top 10 words with the largest changes.

Table 6. Top 10 Words with the Largest Change in the Number of Students Using Them for Question 1 (Patient Explanation) from Pre to Post test

Words with Increased Usage	Change	Words with Decreased Usage	Change
Learning	18	Symptoms	-11
Input	17	Examination	-11
Information	15	Explanation	-9
Bias	15	Presentation	-8
Utilization	13	During	-7
Case	11	Consultation	-6
Convenient	9	View	-6
Judgment	7	Result	-5
Hospital	7	Certain	-5
Misdiagnosis	7	Many	-5

Subsequently, we assessed the structural changes in students' thinking by comparing the cooccurrence networks between pre- and post-tests. The patient explanation network's Jaccard coefficient was 0.1156, indicating very low similarity between the two time points. This dramatic shift was driven by a large number of unique pairs at each time point. The network had only 264

shared cooccurrence pairs; however, 1030 unique pairs existed exclusively at pretest, and 989 unique pairs emerged at posttest. These quantitative findings are supported by a qualitative analysis of the network structures. In the pretest, the network was structured around the case's core elements (Figure 1). However, the network became highly complex and specific at posttest, with keywords like "Learning" and "Data" being directly linked to "AI" and new vocabulary, such as "Bias" and "Blind acceptance," explaining the limitations and risks of AI use (Figure 2).

Responses to Question 2

For Question 2, a significant increase in the use of words related to professional and ethical responsibilities was observed. The number of students using key concepts such as "Information" (increase of 25 students), "Personal info" (+14), "Accuracy" (+13), and "Bias" (+10) increased significantly, as well. Contrastingly, the number of students using abstract or cautionary words, such as "Possibility" (decrease of 9 students), "Reference" (−5), and "Keep in mind" (−5), decreased. Table 7 depicts the top 10 words that underwent the most extensive changes.

Table 7. Top 10 Words with the Largest Change in the Number of Students Using Them for Question 2 (Physician Precautions) from Pre to Post test

Words with Increased Usage	Change	Words with Decreased Usage	Change
Information	25	Possibility	-9
AI	14	Reference	-5
Personal info	14	Oneself	-5
Accuracy	13	Keep in mind	-5
Input	11	Presentation	-4
Bias	10	Knowledge	-4
Patient	8	Examination	-4
Assistance	8	Caution	-3
Final	8	Output	-3
View	7	Source	-3

The cooccurrence network for physician precautions showed a significant change, as well. The Jaccard coefficient was 0.1213, confirming the occurrence of a major structural transformation. The low score was driven by a large number of unique pairs. Only 191 pairs were shared, whereas 631 unique pairs were found exclusively at pretest, and 753 new pairs appeared at posttest. The higher number of unique pairs found at post-test compared to pretest suggests that students not only abandoned previously acquired cautionary topics but also actively explored a wider range of new professional and ethical concerns. These quantitative results were corroborated by a qualitative analysis of network structures. The pretest network comprised three separate, unconnected clusters (Figure 3), which transformed into a single, integrated network at posttest, where "AI" and "Information" acted as hubs (Figure 4).

Students' Perceptions of the Program

Among the 118 participants, 76 (response rate: 64.4%) completed an optional post-program survey based on the ARCS model. On a 5-point Likert scale, students rated the program highly for Attention (mean 4.01, SD 0.62), Relevance (mean 4.20, SD 0.71), Confidence (mean 3.78, SD 0.87), and Satisfaction (mean 4.03, SD 0.75). A repeated-measures ANOVA was conducted to compare and reveal a statistically significant difference between the mean scores of the four ARCS components ($F(3, 225) = 13.15, P < 0.001$). Additionally, post-hoc tests with Bonferroni correction revealed that the mean score for Confidence was significantly lower than the scores for Attention, Relevance, and Satisfaction. Furthermore, Relevance was rated significantly higher than Satisfaction.

Discussion

Principal Findings

This study's primary objective was to evaluate the effectiveness of a newly developed PBL program. Further, this study addresses the following research questions:

RQ1: The educational program improves students' objective knowledge of AI. Accordingly, we hypothesized that foundational knowledge would increase after the lecture phase (from pre- to mid-test) and the subsequent PBL phase (from mid- to posttest) would produce a significant additional increase in practical and ethical knowledge.

RQ2: The educational program qualitatively transforms students' thinking regarding AI-related ethical challenges from an abstract understanding to a concrete and structured thought process based on the clinical context.

Regarding RQ1, although a significant improvement in MCQ scores was observed for the program in its entirety, no statistically significant "additional effect" of the PBL phase was observed. Regarding RQ2, text mining results suggested that after the intervention, students' thinking underwent a qualitative transformation, which was characterized by the use of specific and professional terms, such as "bias," "personal information," and "guidelines," and the development of a structured thought process.

Interpretation of Findings

Regarding RQ1, a clear additional effect of PBL was not detected in MCQ scores due to two main reasons: First, the knowledge measured by MCQ was the foundational content that probably had been efficiently acquired during the lecture phase. The second reason is the limitation of the MCQ format itself. As argued by van der Vleuten and Schuwirth in their paper on PBL assessment, traditional standardized knowledge tests cannot adequately assess abilities such as clinical reasoning and metacognition, which are fostered by PBL[28]. Although MCQs are suitable to measure objective knowledge, they may have low sensitivity in capturing changes in the higher-order practical skills fostered by PBL. Indeed, reviews of earlier research on PBL's effectiveness indicate that whereas knowledge acquisition is comparable to or better than traditional lecture-based learning, PBL is more effective in improving higher-order skills, such as communication, problem-solving, self-learning, and metacognitive competencies [29, 30]. Therefore, the reason a statistically

significant additional effect of PBL was not detected in MCQ scores is that PBL's true effect was reflected in the qualitative changes in students' written responses, rather than the multiple-choice questions themselves.

This is clarified by the results for RQ2. The program was revealed to have qualitatively changed students' thinking. Among the responses to Question 1 (patient explanation), the increased use of words related to AI's mechanisms and limitations, such as "learning," "input," and "bias," suggests that students based their explanations of the bases and limitations of AI's judgments on the Transparency Principle (Explainability) and Accountability Principle detailed in the AI Utilization Guideline and AI Business Guideline [19,20]. Similarly, among the responses to Question 2 (physician precautions), the significant increase in the use of words related to ethical and professional responsibilities, such as "personal information," "accuracy," "bias," "evidence," and "guideline," can be interpreted as the result of students' efforts to learn and apply concepts from guidelines, such as the Privacy Principle and the Fairness Principle, which includes attention to bias [19,20]. The cooccurrence network analysis' quantitative results support the inference that the qualitative change in thinking was accompanied by the structuring of knowledge. The cooccurrence network for Question 1 revealed a very low Jaccard coefficient (0.1156) between the pre- and postintervention networks. This indicates the significant transformation of the contextual relationships between words. For 264 common pairs, there were 1030 unique pairs before and 989 unique pairs after the intervention; accordingly, new knowledge and concepts were established by replacing traditional knowledge. Moreover, the Jaccard coefficient for the cooccurrence network of Question 2 was low (0.1213), indicating a similar structural change. In the Question 2 cooccurrence network, whereas "AI's function," "privacy," and "general caution" existed as separate clusters at T1, a more integrated network was formed with the word "information" as a new hub, connecting "AI" and "diagnosis" with important concepts like "accuracy," at T3. The study's quantitative and qualitative findings resonate with McCoy et al.'s argument that medical students must possess

knowledge about AI. They argue that future physicians must be able to not only "use" AI but also "interpret" its results, recognize potential errors and biases, and "explain" the results and processes to patients and other healthcare professionals [12]. In this study, the qualitative improvement in the postintervention descriptions highlight that students' abilities expanded from the "use" perspective to the "interpret and explain" perspective.

One reason for fostering this qualitative transformation of thinking is interpreted from the perspective of the four components of the ARCS model of motivation: First, the clinical case presented as a video at the beginning of the groupwork session effectively captured students' attention (mean rating 4.01, SD 0.62). A systematic review of emotions' role in medical education highlights the significance of emotional experiences in the learning process [31]. The video depicted a medical student facing an ethical dilemma and potentially fostered emotional engagement, thereby enhancing the students' motivation for subsequent learning. Second, their awareness of the challenge involved in responding to a patient who is over-reliant on AI probably enabled students to clearly understand the learning content's relevance to their future practice (mean rating 4.20, SD 0.71). This educational design aligns with the principles of "Authentic learning," which involves the solving of complex, real-world contextual problems to enhance learners' competency in the practical application of knowledge [32]. Further, students perceived the case in this program as a personal and relevant task due to the effective functioning of the elements of authentic learning. On the other hand, the students' confidence in addressing a complex ethical issue resulted in a mean rating of 3.78 (SD 0.87), which was significantly lower than the values for Attention, Relevance, and Satisfaction ($F(3, 225) = 13.15, P < .001$). Therefore, when the program prompted a qualitative transformation in the students' thinking, a single intervention was insufficient for them to acquire complete confidence in such a complex and uncertain topic as AI ethics. This program was limited to considering the explanation's content to the patient and his son and did not provide any opportunities for practical application, such as a role-play of the explanation. This lack of practical experience probably

contributed to the students' low confidence in applying knowledge. An earlier study highlights the importance of maintaining a balance between confidence and humility and adopting a "not-knowing" stance as a key clinical leadership competency aligned with the Japanese healthcare culture [32]. The low confidence scores observed in our study were interpreted as not any failure to learn, but rather the development of the intellectual humility required by a professional, particularly those confronting complex and uncertain topics such as AI ethics.

However, the high ratings for components other than Confidence, particularly Attention and Relevance, comprised a strong motivation for students to clearly understand the subject's importance and engage in self-directed learning [22]. The effective functioning of these elements enabled the students to complete a learning process involving deep thought, and the resulting insights probably led to the final Satisfaction (mean rating: 4.03, SD 0.75) and structuring of thought observed in this study.

Research Significance and Limitations and Future Research Scope

The primary significance of this study is that it develops, implements, and comprehensively evaluates an educational program themed on a realistic and novel clinical challenge that may be faced by physicians in this era of widespread AI use: responding to a patient who relies heavily on AI. Presenting this ethical dilemma as a video teaching material has pedagogical validity because a systematic literature review indicates that video materials can effectively stimulate student discussion in ethics education [33]. Furthermore, the visualization of the program's educational effect as not only a change in objective knowledge but also a qualitative and structural change in thinking using text mining is a novel approach to evaluate the effects of constructivist learning on complex topics such as AI ethics.

This study has several limitations: First, because it was conducted at a single university, the generalizability of its results may be limited. To address this limitation and enhance the program's generalizability, a multi-institutional collaborative study is warranted. Second, this study considered

only the program's short-term effects on learning. Therefore, the participating students must be subjected to a follow-up study when they proceed to their clinical clerkships to verify the learning effects' persistence. Clarifying how the qualitative transformation in thinking indicated by this study influences students' decision-making and behavior in clinical settings is important to evaluate the value of this type of educational programs. Third, as mentioned earlier, MCQ may not fully capture the effects of PBL. Therefore, there is an urgent need to develop and validate new assessment measures that can objectively assess the practical reasoning and communication skills that cannot be captured by applying MCQs alone.

Conclusions

The PBL program developed and implemented in this study influenced both medical students' knowledge of AI and their thoughts on its ethical challenges. Although the program in its entirety resulted in a significant increase in objective knowledge, a statistically significant additional effect was not confirmed from the PBL phase alone. However, there was a clear change in the qualitative data revealed by text mining: Students' thinking underwent qualitative transformation from an abstract level to a concrete and structured thought process based on the clinical context. Therefore, this study suggests that PBL is an effective pedagogical method to foster higher-order thinking skills in complex topics, such as AI ethics. Hence, this program's design and evaluation framework can serve as a practical model to develop and assess future AI-related curricula in medical education.

Acknowledgments

We extend our sincere gratitude to all the students who participated in this study. Our heartfelt thanks are also directed to Ms. Akane Sumi for her cooperation in the video recording and to Ms. Kazumi Kato for her professional administrative support at the graduate school.

Funding Statement

This work was supported by research funding at the Medical Education Center of Jichi Medical University.

Conflicts of Interest

None declared.

Data Availability

The datasets generated and/or analysed during the current study are not publicly available due to ethical constraints but are available from the corresponding author on reasonable request.

Authors' Contributions

YO was involved in all study aspects. YA contributed to the administration of the educational program and the analysis and interpretation of data. SK contributed to the creation of the evaluation instruments and the analysis and interpretation of data. YM contributed to the conceptualization and supervision of all study aspects and critically reviewed the manuscript. TK and MY contributed significantly to the administration of PBL sessions and the writing of the manuscript. All the authors reviewed and approved the final manuscript.

Abbreviations

AI: artificial intelligence

PBL: problem-based learning

MCQ: multiple-choice questionnaire

IST: information science and technology

MEXT: Ministry of Education, Culture, Sports, Science and Technology

References

1. Masters K. Artificial intelligence in medical education. *Med Teach*. 2019;41(9):976-980. doi:10.1080/0142159X.2019.1595557
2. Paranjape K, Schinkel M, Panday RN, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. *JMIR Med Educ*. 2019;5(2):e16048. doi:10.2196/16048
3. Rincón EHH, Jimenez D, Aguilar LAC, [Flórez](#) JMP, [Tapia](#) AER, [Peñuela](#) CLJ. Mapping the use of artificial intelligence in medical education: a scoping review. *BMC Med Educ*. 2024;24(1):526. doi:10.1186/s12909-025-07089-8
4. Car J, Ong QC, Fox TE, et al. The digital health competencies in medical education framework: an international consensus statement based on a Delphi study. *JAMA Netw Open*. 2024;7(5):e2413131. doi:10.1001/jamanetworkopen.2024.53131
5. Longhini J, Rossetini G, Palese A. Digital health competencies among health care professionals: systematic review. *J Med Internet Res*. 2022;24(8):e36414. doi:10.2196/36414
6. Lee J, Wu AS, Li D, Kulasegaram KM. Artificial intelligence in undergraduate medical education: a scoping review. *Acad Med*. 2021;96(11S):S62-S70. doi:10.1097/ACM.0000000000004291
7. Amiri H, Peiravi S, Shojaee SSR, et al. Medical, dental, and nursing students' attitudes and knowledge towards artificial intelligence: a systematic review and meta-analysis. *BMC Med Educ*. 2024;24(1):412. doi:10.1186/s12909-024-05406-1
8. Çalışkan SA, Demir K, Karaca O. Artificial intelligence in medical education curriculum: an e-Delphi study for competencies. *PLoS One*. 2022;17(7):e0271872. doi:10.1371/journal.pone.0271872
9. Weidener L, Fischer M. Artificial intelligence in medicine: cross-sectional study among medical students on application, education, and ethical aspects. *JMIR Med Educ*. 2024;10:e51247. doi:10.2196/51247

10. Hale J, Alexander S, Wright ST, Gilliland K. Generative AI in undergraduate medical education: a rapid review. *J Med Educ Curric Dev.* 2024;11:23821205241266697. doi:10.1177/23821205241266697
11. Singla R, Pupic N, Ghaffarizadeh SA, Kim C, Hu R, Forster BB. Developing a Canadian artificial intelligence medical curriculum using a Delphi study. *NPJ Digit Med.* 2024;7(1):108. doi:10.1038/s41746-024-01307-1
12. McCoy LG, Nagaraj S, Morgado F, Harish V, Das S, Celi LA. What do medical students actually need to know about artificial intelligence? *NPJ Digit Med.* 2020;3:86. doi:10.1038/s41746-020-0294-7
13. Weidener L, Fischer M. Proposing a principle-based approach for teaching AI ethics in medical education. *JMIR Med Educ.* 2024;10:e55368. doi:10.2196/55368
14. Medical Education Model Core Curriculum Expert Research Committee. The model core curriculum for medical education in Japan 2022 revision. Ministry of Education, Culture, Sports, Science and Technology. https://www.mext.go.jp/content/20250411-mxt_igaku-000028108_00003-2.pdf. Accessed September 20 2025.
15. Onoue T, Asada Y, Imafuku R, et al. Developing competencies relating to information science and technology in Japanese undergraduate medical education. *Med Teach.* 2024;46(sup1):S31-S37. doi:10.1080/0142159X.2024.2385199
16. Ota Y, Asada Y, Mieno M, Matsuyama Y. Competencies required to make use of information science and technology among Japanese medical students: a cross-sectional study. *BMC Med Educ.* 2024;24(1):840. doi:10.1186/s12909-024-05786-4
17. Ota Y, Asada Y, Matsuyama Y. IT: [Proposal of a rubric for 'Competency in Utilizing IT: Information Science and Technology' education]. *Igaku Kyoiku.* 2025;56(1):40-41. doi:10.11307/mededjapan.56.1_40
18. Taylor D, Mifflin B. Problem-based learning: where are we now? *Med Teach.* 2008;30(8):742-

763. doi:10.1080/01421590802217199

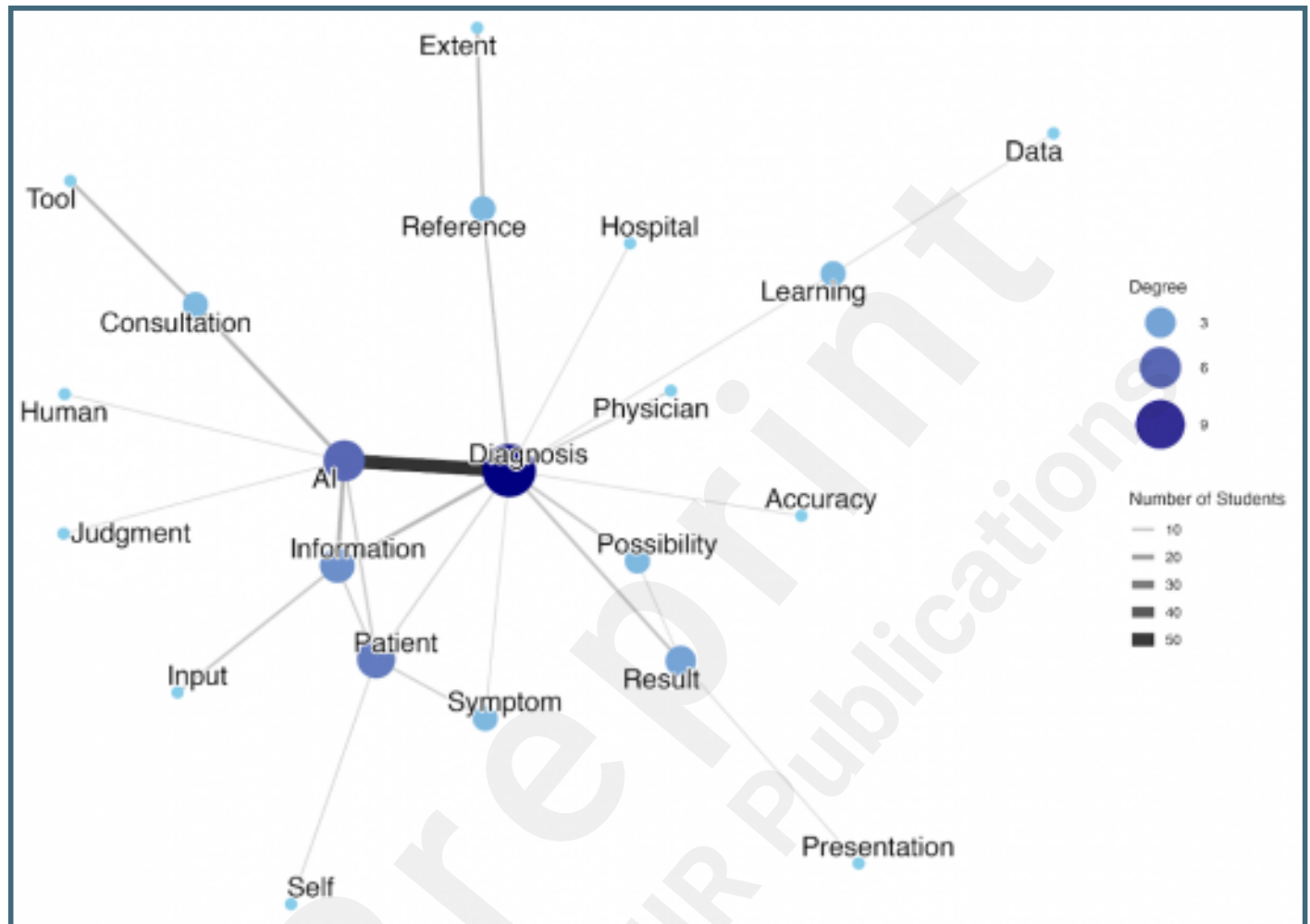
19. The Conference toward AI Network Society. AI utilization guidelines: practical reference for AI utilization. Ministry of Internal Affairs and Communications. 2019.
https://www.soumu.go.jp/main_content/000658284.pdf. Accessed September 20 2025.
20. Ministry of Internal Affairs and Communications, Ministry of Economy, Trade and Industry. AI guidelines for business (Ver. 1.1). 2024.
https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20240419_14.pdf. Accessed September 20 2025.
21. Ota Y, Asada Y, Matsuyama Y. 信頼性向上のためのPBL実践 - ARCSモデルを用いた臨床判断の信頼性向上 - [Implementation of a PBL program on the credibility of AI-based clinical judgments: insights from the ARCS model]. Igaku Kyoiku. 2025;56(3):181-188.
doi:10.11307/mededjapan.56.3_181
22. Keller JM. Motivational Design for Learning and Performance: The ARCS Model Approach. Springer; 2010. doi:10.1007/978-1-4419-1250-3
23. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing; 2023. <https://www.R-project.org/>
24. Silge J, Robinson D. tidytext: text mining and analysis using tidy data principles in R. J Open Source Softw. 2016;1(3):37. doi:10.21105/joss.00037
25. Ishida M. RMeCab: an R interface to MeCab, a Japanese morphological analysis engine. 2008.
<https://github.com/IshidaMotohiro/RMeCab>
26. Kang KA, Han SJ, Chun J, Kim HY. Research trends related to childhood and adolescent cancer survivors in South Korea using word co-occurrence network analysis. Child Health Nurs Res. 2021;27(3):201-210. doi:10.4094/chnr.2021.27.3.201
27. van der Vleuten CPM, Schuwirth LWT. Assessment in the context of problem-based learning. Adv Health Sci Educ Theory Pract. 2019 Dec;24(5):903-914. doi:10.1007/s10459-019-09909-1

28. Trullàs JC, Blay C, Sarri E, Pujol R. Effectiveness of problem-based learning methodology in undergraduate medical education: a scoping review. *BMC Med Educ.* 2022;22(1):104. doi:10.1186/s12909-022-03154-8
29. Stentoft D. Problem-based projects in medical education: extending PBL practices and broadening learning perspectives. *Adv Health Sci Educ Theory Pract.* 2019;24(5):959-969. doi:10.1007/s10459-019-09917-1
30. Lee J, Campbell S, Choi M, Bae J. Authentic learning in healthcare education: a systematic review. *Nurse Educ Today.* 2022;119:105596. doi:10.1016/j.nedt.2022.105596
31. Shikama Y, Oikawa S, Stanyon M, Yasuda M, Otani K. Culturally-aligned clinical leadership competencies for effective teamwork in Japanese healthcare. *BMC Med Educ.* 2024;24(1):1315. doi:10.1186/s12909-024-06272-7
32. Rattani A, Kaakour D, Syed RH, Kaakour AH. Changing the channel on medical ethics education: systematic review and qualitative analysis of didactic-icebreakers in medical ethics and professionalism teaching. *Monash Bioeth Rev.* 2021;39(1):125-140. doi:10.1007/s40592-020-00120-2

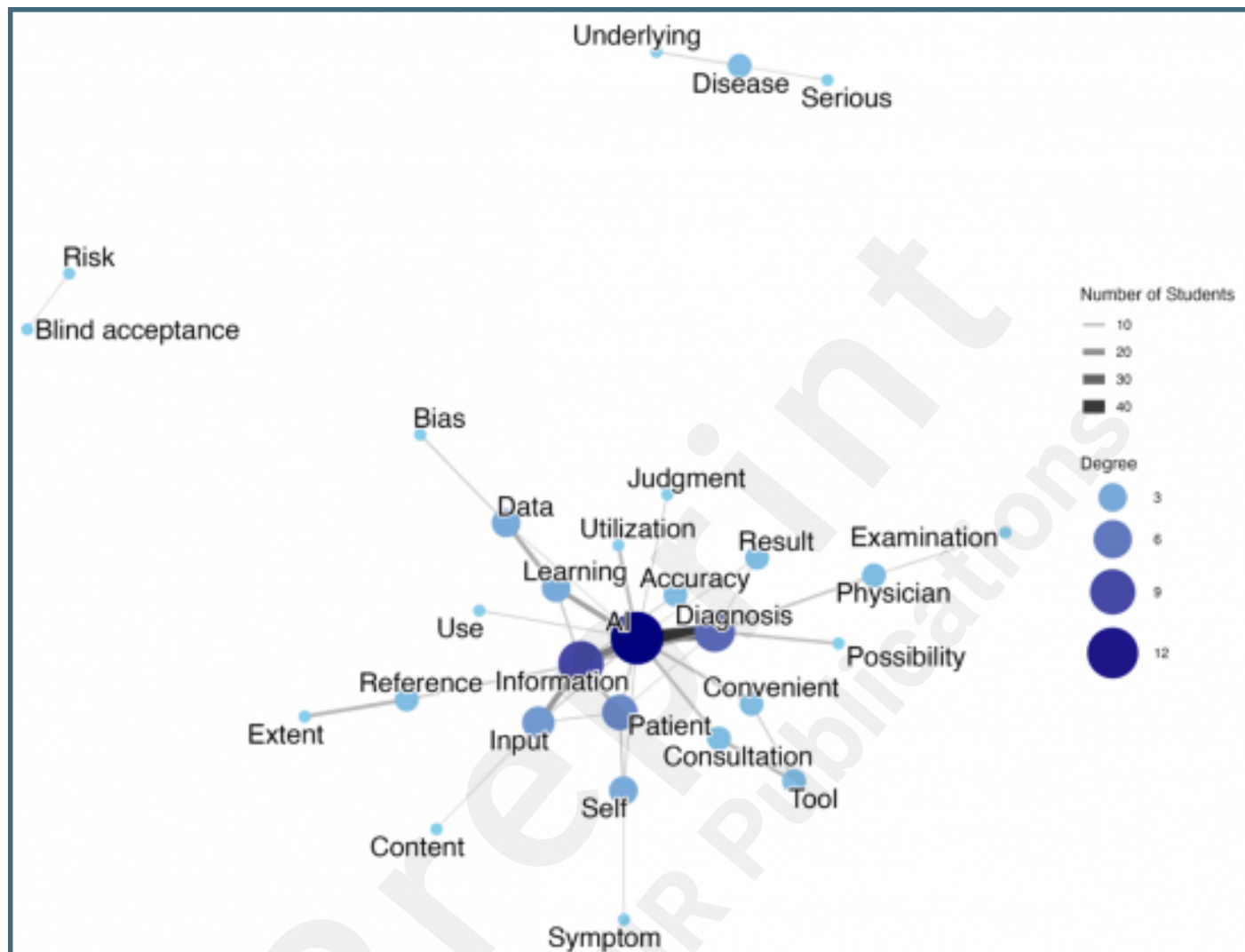
Supplementary Files

Figures

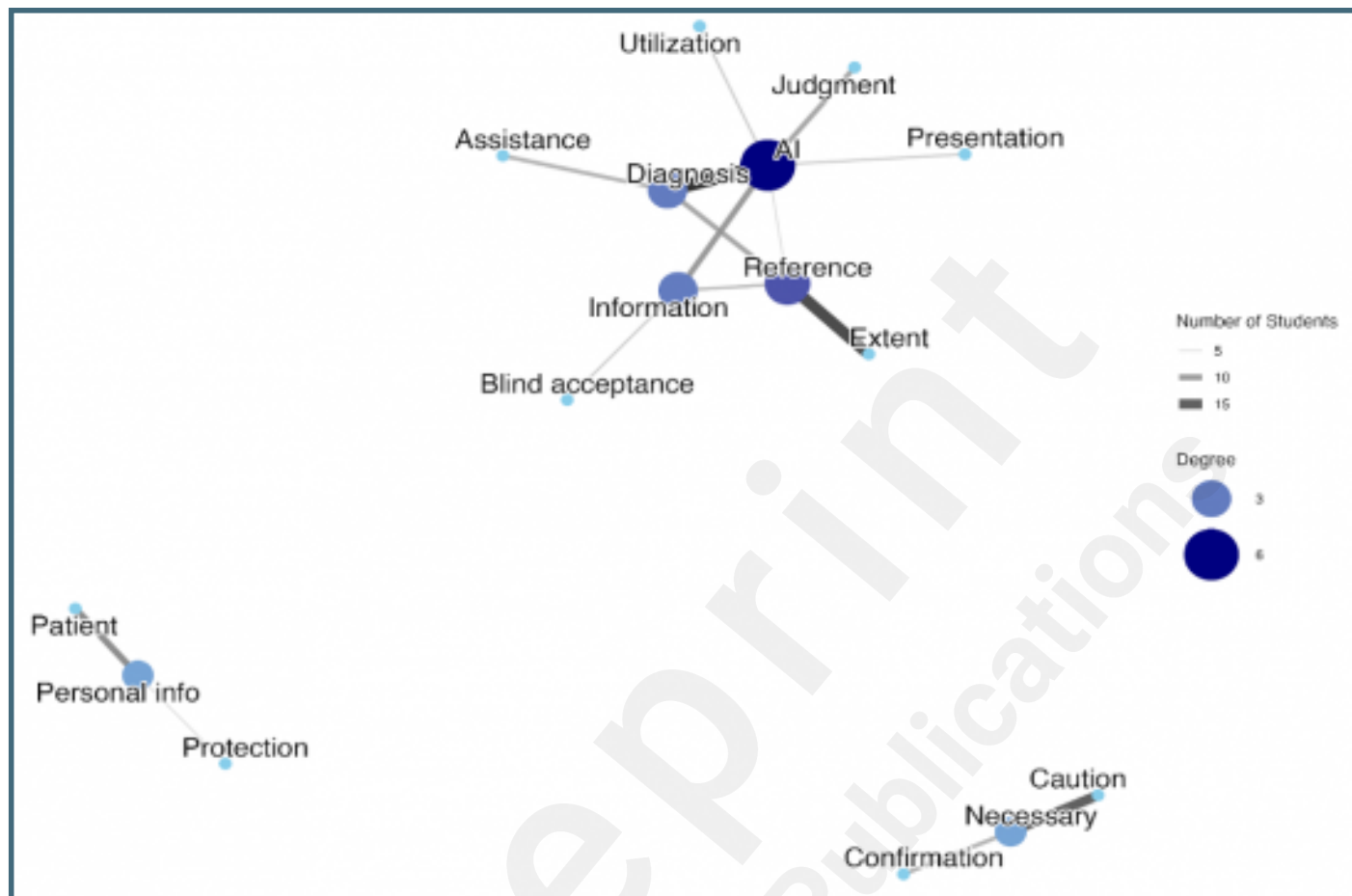
The network visualizes the relationships between keywords in students' explanations to the patient before the program. Each node represents a word, and the size and color of the node indicate its degree (number of connections). The thickness and color of the edge (line) connecting two nodes represent the number of students who used those two words adjacently.



The network visualizes the relationships between keywords in students' explanations to the patient after the program.



The network visualizes the relationships between keywords regarding precautions for physicians before the program.



The network visualizes the relationships between keywords regarding precautions for physicians after the program.

