Advances in Medical Education and Practice

Dovepress
Taylor & Francis Group

Open Access Full Text Article

REVIEW

Thematic Evolution in Medical Education Research (2000–2024): A Large-Scale BERTopic Analysis

Chujie Chen[1,*], Jing Li[2,*], Zhen Zhang[3,*], Tao Zhang[4]

[1]Department of Urology, The Seventh Affiliated Hospital, Sun Yat-Sen University, Shenzhen, 518107, People's Republic of China; [2]Department of Pulmonology, Shenzhen Hospital, Beijing University of Chinese Medicine, Shenzhen, 518100, People's Republic of China; [3]Department of Endocrinology, The Seventh Affiliated Hospital, Sun Yat-Sen University, Shenzhen, 518107, People's Republic of China; [4]Department of Urology, Shenzhen Hospital, Beijing University of Chinese Medicine, Shenzhen, Guangdong, 518100, People's Republic of China

*These authors contributed equally to this work

Correspondence: Tao Zhang, Department of Urology, Shenzhen Hospital, Beijing University of Chinese Medicine, Shenzhen, Guangdong, 518100, People's Republic of China, Email zhangtaomed@163.com

Background: Few large-scale, semantically informed mappings of medical education research exist: traditional methods such as keyword co-occurrence and Latent Dirichlet Allocation cannot capture the contextual meaning of scientific abstracts. This study aimed to identify major research topics in medical education, classify their temporal trends, explore temporal associations with major policy and technology developments, and characterize their structural positioning.

Methods: We applied BERTopic, a transformer-based topic modeling framework, to 276,253 English-language publications from PubMed, Web of Science, Scopus, and OpenAlex (2000–2024). Documents were embedded with a sentence-transformer language model and clustered into topics; labels were independently reviewed by two authors. Trends in each topic's relative share were classified using Poisson regression with an offset for total annual output, autocorrelation-robust Newey–West standard errors, and false discovery rate correction. Topic positioning used an embedding-based centrality-density heuristic.

Results: BERTopic identified 39 topics (7.0% outliers); Topic 0 was a residual, non-substantive cluster excluded from interpretation. Among the 38 substantive topics, 10 showed significantly increasing relative shares (Hot; including AI/ChatGPT, virtual reality simulation, burnout and wellbeing, diversity and equity, empathy, and climate and health education), 12 showed decreasing shares (Cold), and 16 were Stable. Classifications agreed across alternative standard-error specifications for 35 of 38 topics; simulation-based education, clinical imaging, and interprofessional education were significant only under less conservative errors. Temporal associations with policy and technology reference periods were exploratory and not causal. The positional map distinguished established themes from emerging, terminologically unsettled niches.

Conclusion: The relative emphasis of medical education research has gradually shifted toward technology-enhanced and socially responsive themes, while foundational topics continue to grow in absolute volume. These findings offer one data-driven, assumption-dependent lens on the field's evolution to inform curriculum planning, research prioritization, and policy.

Keywords: natural language processing, machine learning, text mining, bibliometrics, research trends, health professions education

Introduction

Medical education research has grown substantially over the past quarter century: annual publication volume increased from approximately 3,500 articles in 2000 to over 17,000 by 2024. This expansion has been accompanied by a diversification of research themes, driven by evolving pedagogical frameworks, technological innovations, and shifting societal expectations of health professions education.[1] Understanding how these research themes have emerged, evolved, and declined over time is essential for identifying knowledge gaps and for guiding funding priorities and strategic planning in the field.

Bibliometric methods have been increasingly applied to map the intellectual structure of medical education.[2,3] Traditional approaches have relied on keyword co-occurrence analysis, co-citation networks, and Latent Dirichlet Allocation (LDA) topic modeling to identify research clusters and temporal trends. However, these methods share

https://doi.org/10.2147/AMEP.S618140
Received: 19 April 2026
Accepted: 23 June 2026
Published: 30 June 2026

important limitations: keyword-based approaches depend on author-assigned or indexed terms that may not capture the full semantic content of publications, while LDA treats documents as bags of words without considering word order or contextual meaning, limiting its ability to disambiguate polysemous terms common in interdisciplinary fields such as medical education.[4] These limitations are especially consequential in medical education, where closely related constructs are described with different vocabularies across clinical specialties, educational paradigms, and social science traditions; methods that depend on exact term matching can therefore fragment a single theme into several keyword clusters, or conflate distinct themes that share surface vocabulary.

Recent advances in natural language processing have introduced transformer-based topic modeling approaches that overcome these limitations. BERTopic, a modular framework combining sentence-transformer embeddings with dimensionality reduction and density-based clustering,[5,6] has been adopted in bibliometric studies across diverse fields, including public health,[7] environmental science,[8] and information systems.[9] Because it uses pre-trained language models to generate contextual document embeddings, BERTopic captures semantic relationships that bag-of-words approaches cannot, making it particularly suitable for short scientific texts such as abstracts. Among embedding-based alternatives, BERTopic was preferred over Top2Vec[10] and dynamic topic models because its modular pipeline scales to corpora of this size, produces transparent keyword-based topic representations that human reviewers can inspect, and has the most extensive validation record across disciplines.[7–9] Transformer-based topic models nevertheless pose interpretability challenges: topics are induced from high-dimensional embeddings that are not directly human-readable, so topic labels require structured human review and should be treated as convenient summaries rather than definitive categories.[9] Large-scale machine-learning analyses in adjacent fields, such as semantic mappings of the COVID-19 research corpus,[11] have demonstrated the feasibility of this approach for corpora exceeding 100,000 documents. However, few studies have applied BERTopic at the scale of a full multi-database medical education corpus with integrated temporal, event-level, and structural analyses. These methodological developments are also unfolding within a broader transformation of scholarly communication, in which artificial intelligence increasingly mediates how literature is searched, synthesized, and reviewed;[12] large-scale semantic mappings such as the present study should be read against this changing publishing environment.

Existing bibliometric analyses of medical education have been limited in scope, typically focusing on specific subdisciplines, individual journals, or citation-based rankings,[3,13,14] and generally analyzing fewer than 10,000 records using traditional keyword-based or LDA methods. Beyond quantifying publication growth, mapping a field's thematic structure and positioning has theoretical value in its own right: it reveals which themes anchor the field's intellectual core, which operate as specialized niches, and which are emerging or receding – information that no single study or journal-level review can provide and that can support curriculum renewal, research prioritization, and policy planning.[2] The aim of this study is therefore not to catalogue publication counts but to characterize how the thematic organization of medical education scholarship has evolved. An analysis that spans all subdisciplines, all major databases, and the complete 2000–2024 timeframe provides a more complete basis for that characterization than subdiscipline- or journal-level studies.

The objectives of this study were to: (1) identify and characterize the major research topics in medical education from 2000 to 2024 using BERTopic on a corpus of 276,253 publications; (2) classify temporal trends in each topic's relative share as Hot (increasing), Cold (decreasing), or Stable using Poisson generalized linear models with autocorrelation-robust standard errors and false discovery rate correction; (3) explore the temporal association between major policy and technology events and topic-level publication rates using exploratory interrupted time series analysis; and (4) characterize the lifecycle stage and structural positioning of each topic using S-curve modeling and an embedding-based adaptation of the strategic diagram framework.

Materials and Methods

Data Source and Corpus Construction

Publication records were retrieved from a consolidated multi-database bibliometric dataset integrating PubMed (via MEDLINE), Web of Science Core Collection, Scopus, and OpenAlex, covering the medical education literature from 2000 to 2024 (search dates: January 15–February 20, 2026). The search strategy used medical education-related MeSH terms, subject categories, and keyword combinations applied to each database according to its indexing structure; full per-database Boolean queries and raw per-database hit counts are provided in Supplementary Table S1b, with the pre-specified search window and filters in

Supplementary Table S1a, the cross-database deduplication and filtering steps in Supplementary Table S1c, and the per-source raw retrieval counts in Supplementary Table S1d; the per-database contribution to the final analytical corpus is summarized in Supplementary Table S2.

Raw retrieval across the four databases produced 943,925 records; cross-source deduplication by DOI > PMID > normalized-title similarity reduced this pool to 622,407 unique records, from which the final corpus was built through three sequential filters: (1) exclusion of 282,061 non-English publications; (2) exclusion of 56,489 records without abstracts; and (3) exclusion of 7,604 records with abstracts below 50 words after text cleaning. Title similarity used a Jaccard coefficient computed over case-folded word tri-grams (threshold ≥0.90), combined with matching first-author surname and publication year to guard against false positives on generic titles. Partial 2025 records (a small residual remaining from indexing overlap between the four sources' most recent update windows) were retained in the retrieved dataset for corpus completeness but excluded from the primary 2000–2024 analysis; they appear only in the ITS sensitivity check reported in Supplementary Table S3. Text preprocessing removed HTML tags, structured section labels (eg., "Background:", "Methods:"), URLs, Email addresses, and Unicode artifacts. The final analytical corpus comprised 276,253 English-language publications with cleaned abstracts. The corpus construction flow is summarized below:

Corpus Construction Flow: - Raw retrieval yielded 943,925 records from PubMed, Web of Science, Scopus, and OpenAlex. Cross-source deduplication using DOI, PMID, and title tri-gram Jaccard similarity with author/year matching reduced the dataset to 622,407 unique records. Sequential exclusions removed 282,061 non-English publications, 56,489 records without abstracts, and 7,604 records with cleaned abstracts shorter than 50 words. The final analytical corpus comprised 276,253 publications for the primary 2000–2024 analysis; partial 2025 records were retained only for sensitivity analyses.

The search strategy was designed to maximize sensitivity for medical education research indexed in biomedical and multidisciplinary databases. The four databases were selected for their complementary coverage: PubMed provides MeSH-indexed biomedical literature, Web of Science and Scopus capture interdisciplinary journals not indexed in MEDLINE, and OpenAlex provides open-access metadata with topic-level classification. Search terms were adapted to each database's indexing structure (MeSH terms for PubMed, free-text for WoS/Scopus, topic IDs plus free-text for OpenAlex), which introduces some asymmetry but reflects standard multi-database search practice. Education-specific databases (eg., ERIC) were not included, which may underrepresent publications in education journals not indexed in the four selected databases. The resulting corpus should therefore be interpreted as a large English-language multi-database medical education corpus rather than an exhaustive representation of the entire field.

Embedding Generation

Document embeddings were generated using the all-MiniLM-L6-v2 sentence-transformer model,[15] which maps text inputs to 384-dimensional dense vector representations optimized for semantic similarity tasks. This general-purpose model was selected over domain-specific alternatives (eg., PubMedBERT, BioBERT) for three reasons: (1) it is optimized for short-text semantic similarity, which is the primary task in abstract-level topic modeling; (2) it has been extensively benchmarked in BERTopic applications across diverse fields, facilitating methodological comparability; and (3) medical education abstracts contain a mix of biomedical, educational, and social science terminology that may not benefit as strongly from biomedical-only pretraining as clinical texts would. Each abstract was encoded into a 384-dimensional embedding, yielding a 276,253 × 384 document-embedding matrix. A stratified random subsample of 50,000 documents (stratified by publication year) was reserved as a pilot set for parameter tuning.

Topic Modeling with BERTopic

Topics were identified using BERTopic (version 0.17.4), a modular framework that combines transformer-based document embeddings with dimensionality reduction and density-based clustering. The pipeline consisted of four stages: (1) dimensionality reduction via UMAP[16] to five dimensions for clustering (cosine metric, n_neighbors = 15), with a separate two-dimensional projection computed solely for the topic-landscape visualization presented in the Results; (2) density-based clustering via HDBSCAN[17] (min_cluster_size = 200); (3) topic-representation extraction via class-based TF-IDF (c-TF-IDF) over unigrams and bigrams with the default English stop-word list; and (4) fine-tuned topic labeling via KeyBERTInspired and Maximal Marginal Relevance[18] representation models, each producing the top 10 keywords per topic. A single random seed

(42) was applied consistently across UMAP, HDBSCAN, BERTopic's agglomerative topic reduction, and the cosine-similarity outlier reassignment step, so that the final topic structure is fully reproducible given the same input corpus and software versions; the complete hyperparameter configuration is specified in the analysis code. The KeyBERTInspired and MMR layers re-rank keywords by embedding-space semantic relevance, reducing the influence of any residual high-frequency functional terms in the final topic representations used for labeling and interpretation.

Parameter Tuning and Model Selection

Hyperparameters were optimized on the 50,000-document pilot set using a two-stage protocol. In the first stage (T1), a grid search evaluated three HDBSCAN min_cluster_size values (100, 200, 500) crossed with five target topic counts (40, 50, 60, 70, 80) applied via BERTopic's reduce_topics function. Each configuration was scored by c_v topic coherence (computed via gensim on a 10,000-document subsample, processes = 1) plus topic diversity, with an outlier proportion threshold of 20%. The configuration maximizing the composite score while satisfying the outlier constraint was selected (best: min_cluster_size = 200, nr_topics = 40).

In the second stage (T2), model stability was assessed by fitting the selected configuration across five random seeds (42, 123, 456, 789, 1024), applying reduce_topics to each, and computing pairwise Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) on non-outlier documents only. The mean ARI of 0.787 (SD = 0.048) exceeded the prespecified stability threshold of 0.70, confirming acceptable reproducibility.

Final Model Fitting and Outlier Reduction

The final model was fitted on the full 276,253-document corpus. The raw model yielded 241 initial clusters, which were reduced to 40 topics via BERTopic's agglomerative topic reduction. Outlier documents (initially 32.7% of the corpus) were reassigned to the nearest topic cluster using cosine similarity in the original 384-dimensional embedding space (threshold = 0.5), followed by topic representation updating. The reassignment step merged one of the 40 topics into a neighboring cluster (the only topic whose final size fell below the min_cluster_size threshold after outlier absorption was absorbed into Topic 0), yielding 39 final topics covering 256,865 documents and reducing the outlier proportion to 7.0%. Throughout the manuscript, "39 topics" refers to this post-reassignment substantive set (with Topic 0 retained as a residual/background cluster).

Temporal Trend Analysis

Topic-level temporal trends were analyzed using annual document counts for each of the 39 topics across the 25 annual observations of 2000–2024. Because the total volume of the corpus grew from roughly 3,500 to more than 17,000 publications per year, the absolute count of nearly every topic rises over time; trend analysis must therefore separate topic-specific dynamics from this background growth. To do so, Poisson generalized linear models (GLMs) were fitted with year as the predictor and the logarithm of total annual non-outlier publications as an offset term. The offset converts each model from one of absolute counts to one of relative share: the resulting incidence rate ratio (IRR) per year describes how a topic's proportion of the field changes annually (an IRR of 1.05 means the topic's share grows by about 5% per year relative to the field as a whole), so secular corpus growth is balanced out by construction.

Annual publication series violate two assumptions of the basic Poisson model. First, all large topics showed overdispersion (deviance-to-degrees-of-freedom ratio > 1.5 in 36 of 39 models). Second, residual diagnostics revealed substantial first-order serial autocorrelation: Durbin–Watson statistics fell well below 2 for most large topics (eg., 0.45 for Topic 1) and Ljung-Box tests were significant (Supplementary Table S4). Primary inference therefore used Newey–West heteroskedasticity- and autocorrelation-consistent (HAC) standard errors (Bartlett kernel, maxlag = 3), which are robust to both violations. P-values were corrected for multiple testing across all 39 topics using the Benjamini–Hochberg false discovery rate (FDR) procedure at alpha = 0.05, and topics were classified as Hot (FDR-adjusted $P < 0.05$, positive beta), Cold (FDR-adjusted $P < 0.05$, negative beta), or Stable (FDR-adjusted $P \geq 0.05$). Two sensitivity analyses accompanied this primary specification: Huber-White sandwich (HC1) robust standard errors, which address overdispersion but not serial dependence and are therefore less conservative; and nonparametric Mann-Kendall trend tests with tie correction. Agreement between the HAC and HC1 classifications, including the topics whose classification depends on the choice of error structure, is reported explicitly in the Results.

Human-interpreted topic labels were assigned by two of the study authors through a structured two-stage independent review. In the first stage, each author independently inspected, for every topic, (i) the top-10 c-TF-IDF keywords, (ii) the top-10 KeyBERT-refined keywords, and (iii) the top-5 representative documents (full titles and abstracts) automatically selected by BERTopic on the basis of c-TF-IDF cosine proximity to the topic vector. Each author proposed an initial label without seeing the other's choice. In the second stage, the two label sets were compared: 32 of the 39 topics (82.1%) received identical or near-synonymous labels in independent review (eg., "AI and ChatGPT" vs "Generative AI in medical education"); the remaining 7 were reconciled through joint discussion of keyword and document evidence, with no topic requiring more than a single discussion round. The full top-5 representative titles per topic are provided in Supplementary Table S5 to allow readers to inspect the empirical basis of each label. We did not engage an external expert panel for blinded label review; this is a limitation of the topic validation procedure, and external multi-rater validation would be required before any individual label is treated as a definitive characterization. Topic coherence ($c_v = 0.512$) and topic diversity (0.382) were reported during model selection as quantitative complements to the qualitative label review. A topic-label confidence heuristic computed as the lexical overlap between each topic's human label and its top-5 representative titles (Supplementary Table S6) provides an additional descriptive check but should not be over-interpreted: a "low-overlap" outcome indicates that the human label uses abstract vocabulary (eg., "pedagogy", "IPE") while titles use concrete specialty terminology, not necessarily that the label is wrong; Topic 14 is the one confirmed case of genuine label-document heterogeneity and is discussed in the Limitations section.

Event Correlation via Exploratory Interrupted Time Series

The temporal association between five core events in medical education and topic-level publication rates was explored using interrupted time series (ITS) Poisson GLMs in an exploratory, descriptive framework. These results should be interpreted as exploratory signals of temporal coincidence rather than evidence of causal effects, given the limited number of annual data points, serial autocorrelation not explicitly modeled, and the overlapping post-event windows of the 2020–2022 events. The five events were: WFME Global Standards (2003),[19] ACGME Milestones (2013),[20] COVID-19 Pandemic (2020), USMLE Step 1 Pass/Fail transition (2021), and ChatGPT Release (2022). These reference points were selected a priori as the accreditation, policy, and technology milestones most frequently invoked in the medical education literature, each anchored to a specific calendar year as required by the ITS design. We acknowledge that ACGME Milestones and the USMLE Step 1 transition are primarily North American developments; they were retained because North American journals contribute a large share of the corpus, and their regional character is considered when interpreting the results. Competency-based medical education more broadly was not modeled as a separate reference event because its adoption has been gradual and regionally staggered rather than tied to a single global time point, making it unsuitable for a point-interruption design. Each model included terms for baseline trend (year), level change (binary indicator for post-event years), and slope change (interaction of year with post-event indicator), with an offset for log total annual publications. Where overdispersion was detected (deviance/df > 1.5), HC1 robust standard errors were applied. For the per-event ITS models, p-values for level and slope change coefficients were jointly corrected across all 390 tests (5 events × 39 topics × 2 coefficients) using Benjamini–Hochberg FDR; per-topic per-event results are tabulated in Supplementary Table S7. The compound 2020–2022 ITS model (see Sensitivity and Robustness below) is a separate analysis whose 2-coefficient × 39-topic p-values are FDR-corrected within that model only (independent of the per-event FDR pool) and reported in Supplementary Table S8, so that FDR budgets are not mixed across analyses with different null structures. Pre/post mean frequency comparisons used symmetric two-year windows excluding the event year. Changepoint detection was performed using the PELT algorithm (l2 cost model) with a BIC-derived penalty.

Embedding-Based Positional Mapping

Topic positioning was characterized using two complementary approaches inspired by Callon's co-word strategic diagram framework[21] but adapted to operate on embedding representations rather than traditional co-word networks. The centrality and density metrics capture different structural properties than co-occurrence frequency, and the quadrant assignments should be interpreted as heuristic descriptors rather than direct equivalents of the traditional co-word framework. The embedding-based structural diagram defined centrality as the mean cosine similarity between each topic's 384-dimensional centroid vector and

all other topic centroids (Centrality_i = [1/(K-1)] * sum of cos(c_i, c_j) for all j != i, where K = 39 topics), and density as the mean pairwise cosine similarity among each topic's top-10 keyword embeddings (Density_i = [2/(m*(m-1))] * sum of cos (kw_p, kw_q) for all p < q, where m = 10 keywords). These metrics are global, time-invariant structural properties. The time-varying co-movement diagram computed per-period centrality as the mean absolute Pearson correlation of a topic's annual proportion vector with all other topics' vectors within each time window (2000–2008, 2009–2016, 2017–2024), enabling detection of quadrant migration across periods. Density was held constant (global keyword embeddings) across both approaches. Quadrant assignment (Motor, Niche, Basic/Transversal, Emerging/Declining) used period-specific median centrality and density as thresholds.

Sensitivity and Robustness Analyses

In addition to the HC1 and Mann-Kendall sensitivity checks for the trend classification described above, seven further sensitivity analyses were conducted to assess the robustness of the primary findings. (1) Compound event ITS: To address the temporal clustering of events in 2020–2022 (COVID-19, USMLE Step 1 Pass/Fail, ChatGPT), we modeled this period as a single compound disruption in a combined ITS analysis. (2) S-curve lifecycle analysis (supplementary): Logistic S-curves were fitted to per-topic cumulative proportions to classify lifecycle stages (Emerging/Growing/Mature/Saturated); due to limited discriminatory power, results are reported as supplementary context only and are not discussed substantively in the main text. (3) BERTopic versus LDA comparison: Latent Dirichlet Allocation (k=39, online learning, 20 iterations) was fitted on the same corpus using identical CountVectorizer preprocessing; aggregate quality indices are reported in Supplementary Table S9 and a six-theme qualitative comparison of top-10 terms is shown in Supplementary Figure S1. This comparison is descriptive only and is not used as a basis for claiming superiority of either method, given that diversity, coherence, and assignment-confidence indices favor different methods and the metrics are not directly commensurable across model families. (4) Outlier threshold sensitivity: The cosine similarity threshold for outlier reassignment was varied from 0.3 to 0.7 (primary analysis: 0.5) to assess its effect on outlier proportion and topic structure. (5) Topic 0 exclusion: Trend classification and ITS were re-run excluding the catch-all Topic 0. (6) Partial-year exclusion: All temporal analyses were repeated limiting data to 2000–2024, excluding the partial 2025 data. (7) Quadrant-threshold sensitivity for the embedding-based positional map: Quadrant assignments were recomputed using the mean centrality and density (instead of the median) as splits, to assess how many topics change quadrant label under an alternative threshold definition.

Software and Reproducibility

All analyses were performed using Python 3.13.9 with BERTopic 0.17.4, sentence-transformers 5.2.3, UMAP 0.5.11, HDBSCAN 0.8.41, scikit-learn 1.7.2, statsmodels 0.14.5, SciPy 1.16.3, gensim 4.0.0, ruptures 1.1.10, NumPy 2.3.5, and pandas 2.3.3. A fixed random seed (42) was used throughout for reproducibility. Figures were generated using matplotlib.

Ethical Considerations

This study analyzed publicly available bibliometric metadata from published academic literature. No human subjects were involved, and no individual-level data were collected. Ethical review was therefore not required.

Results

Thematic Landscape

A total of 276,253 publications from the medical education corpus (2000–2024) was analyzed using BERTopic. The model identified 39 distinct research topics after topic reduction and outlier reassignment, with 256,865 documents (93.0%) assigned to a substantive topic and 19,388 (7.0%) remaining as outliers. Publication volume increased steadily across the three study periods, from 49,616 documents in 2000–2008 to 85,992 in 2009–2016 and 121,529 in 2017–2024 (Figure 1). We treat Topic 0 (n = 25,095, 9.8% of non-outliers) throughout the manuscript as a residual/background cluster: it comprises semantically heterogeneous documents whose top keywords are dominated by non-discriminative terms (eg., "the", "medical", "education"), and it does not represent a substantive thematic domain. Two further small clusters are journal-metadata artifacts rather than research themes: Topic 36 (JAMA-family journal section headers and

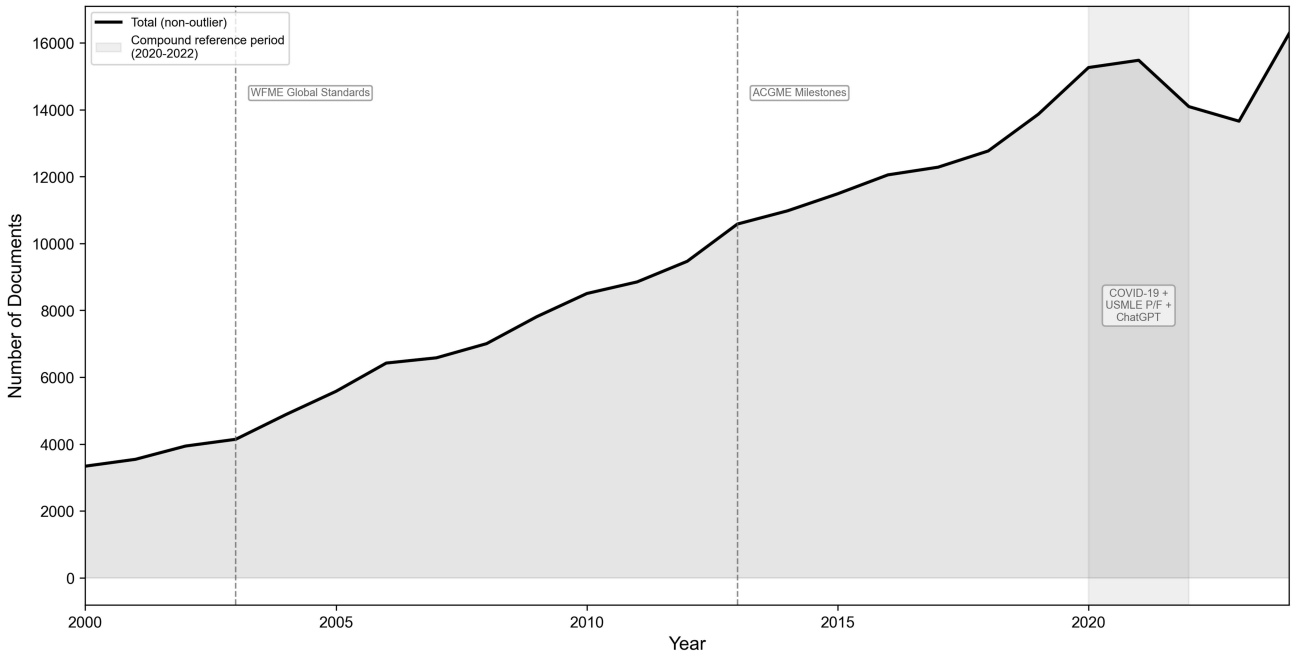

Figure 1 Annual publication volume and historical reference periods (2000–2024). The black curve shows the annual number of non-outlier documents. Two temporally isolated reference points are marked as gray dashed lines: WFME Global Standards (2003) and ACGME Milestones (2013). The shaded gray band marks the 2020–2022 compound reference period (COVID-19, USMLE Step 1 Pass/Fail transition, and ChatGPT release), modeled jointly in the auxiliary interrupted time series analysis due to temporal overlap. These markers indicate temporal context only and should not be interpreted as causal events.

specialty indexing tokens) and Topic 38 (regional Southeast Asian medical-education journal-name tokens). All three clusters are retained in tables for completeness and structural transparency but excluded from substantive interpretation. A sensitivity analysis re-running the trend classification with Topic 0 removed from both the model pool and the offset denominator left all 38 substantive classifications unchanged (Supplementary Table S10), confirming that the findings reported below are not driven by this residual cluster (full stability diagnostics are reported in the Topic Stability and Robustness subsection). A qualitative comparison with LDA is reported in the Sensitivity and Robustness section.

The UMAP two-dimensional projection of document embeddings revealed a complex but structured semantic space (Figure 2). Topics formed recognizable clusters in embedding space, with several discipline-specific topics (dental education, pharmacy education, nursing education) occupying peripheral positions, while more general topics (assessment, professionalism) clustered centrally. The five largest substantive topics were student learning and pedagogy (Topic 2, n = 19,032), surgical skills training (Topic 1, n = 15,981), simulation-based education (Topic 3, n = 15,232), nursing education (Topic 5, n = 14,222), and clinical imaging and ultrasound (Topic 4, n = 12,947), collectively accounting for 30.1% of non-outlier documents (Table 1).

Temporal Trends in Relative Topic Shares

Poisson generalized linear models with an offset for annual publication volume, Newey–West HAC standard errors, and Benjamini–Hochberg false discovery rate correction classified 10 of the 38 substantive topics as Hot (significantly increasing in relative proportion), 12 as Cold (significantly decreasing), and 16 as Stable (Table 2); the residual Topic 0 was classified Cold but is not interpreted. Under the less conservative HC1 sensitivity specification, which corrects overdispersion but not serial autocorrelation, three additional small-effect topics reached significance: simulation-based education (Topic 3, Hot), clinical imaging and ultrasound (Topic 4, Cold), and interprofessional education (Topic 10, Hot). The remaining 35 of 38 substantive classifications were identical (Supplementary Table S4). These three topics should therefore be regarded as having trend evidence that depends on the choice of error structure, and we do not count them among the Hot or Cold findings. Mann-Kendall trend tests provided concordant directions as an additional

https://doi.org/10.2147/AMEP.S618140

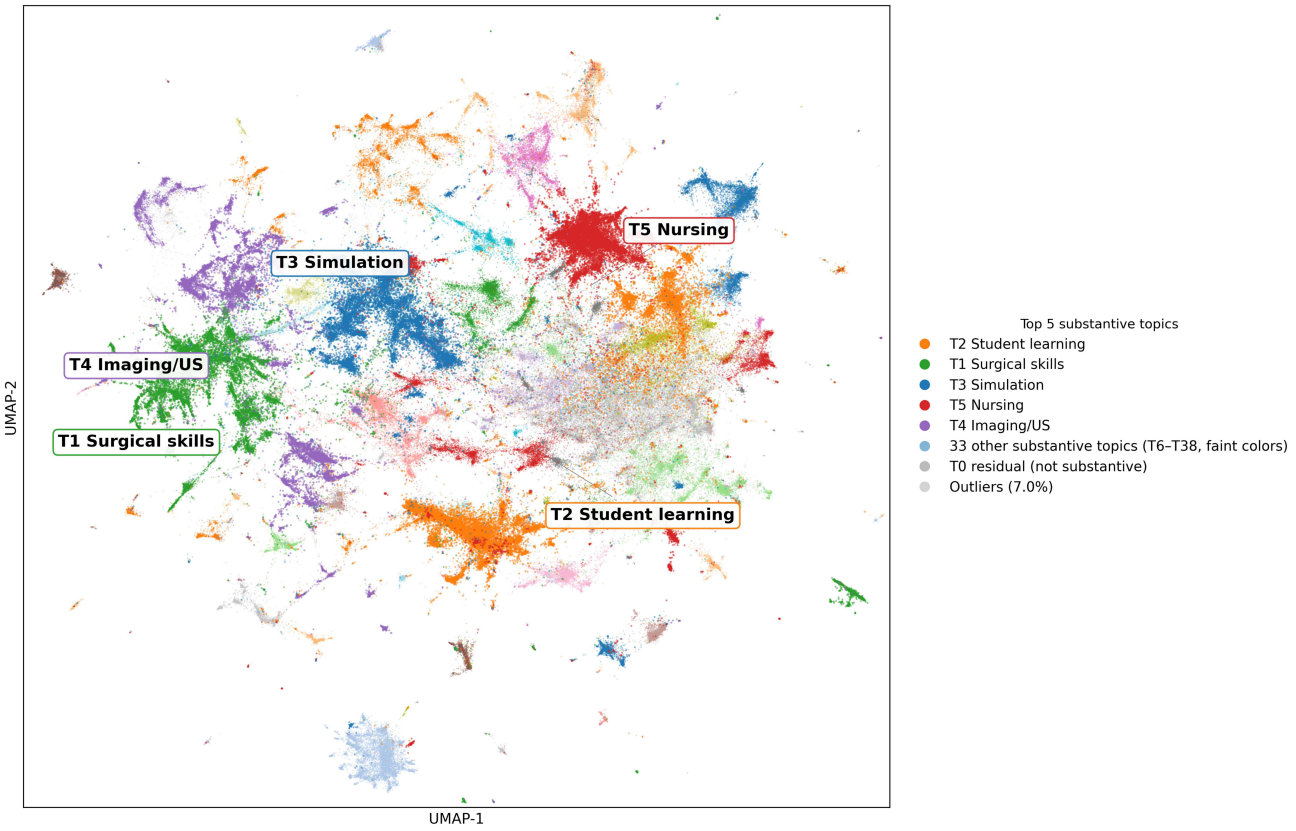

Figure 2 Topic distribution in two-dimensional semantic space. Each point represents one of 276,253 documents projected onto two UMAP dimensions derived from sentence-transformer embeddings (all-MiniLM-L6-v2). The five largest substantive topics (T1 Surgical skills, T2 Student learning, T3 Simulation, T4 Imaging/US, T5 Nursing) are highlighted in saturated colors with bold labels at their centroid positions; the 33 other substantive topics (T6-T38) are rendered in faint distinct colors for visual cluster distinction only; specific topic-color mappings are not preserved at this display scale, and readers should refer to Supplementary Table S5 for the full 39-topic list (with top-10 keywords and representative titles per topic) and Table 1 for the top-15 topic overview. Topic 0 (residual/background cluster) and outlier documents (7.0%) are rendered in gray and light gray respectively to de-emphasize their non-substantive role. Peripheral positions of discipline-specific topics (dental, pharmacy, nursing) reflect lexical and indexing conventions as well as semantic distance, and should not be read as evidence of pedagogical isolation.

robustness check. Full effect sizes (beta coefficients, standard errors, incidence rate ratios with 95% confidence intervals, and both HAC and HC1 p-values) for all 39 topics are reported in Supplementary Table S11.

Hot topics comprised student learning and pedagogy (Topic 2), burnout and wellbeing (Topic 9), diversity and equity (Topic 12), surgical residency programs (Topic 13), vaccination education (Topic 14), AI and ChatGPT in education (Topic 16), anatomy and dissection (Topic 19), virtual reality simulation (Topic 25), empathy (Topic 29), and climate change and health (Topic 34). Cold topics comprised nursing education (Topic 5), dental education (Topic 7), nutrition and public health education (Topic 8), medical professionalism as traditionally defined (Topic 11), palliative care and oncology education (Topic 17), psychiatry and mental health education (Topic 18), airway management training (Topic 23), geriatrics education (Topic 26), duty hour restrictions (Topic 32), complementary and alternative medicine (Topic 33), spirituality and religious care (Topic 35), and tuberculosis education (Topic 37).

The stacked area chart (Figure 3A) and heatmap (Figure 3B) illustrated these temporal dynamics across all 39 topics. The proportional share of student learning and of technology-related topics expanded markedly from 2015 onward, while several traditional topics showed a declining relative share despite absolute growth in publication volume.

An exploratory S-curve lifecycle analysis is provided in the Supplementary Materials (Supplementary Table S12 and Supplementary Figure S2). The analysis classified 27 of 33 fitted topics as Mature (82%), reflecting the mathematical properties of normalized cumulative curves rather than definitive lifecycle staging; it is therefore reported as Supplementary Context rather than a core analytical framework.

Table 1 Overview of the 15 Largest Topics Identified by BERTopic (Full 39-Topic Table in Supplementary Table S5)

Rank	Topic	Interpreted Label	Documents	Trend	Peak Period
1[†]	0	General/Residual cluster (not a substantive topic)	25,095	Cold	2009–2016
2	2	Student learning and pedagogy	19,032	Hot	2017–2024
3	1	Surgical skills training	15,981	Stable	2009–2016
4	3	Simulation-based education	15,232	Stable	2009–2016
5	5	Nursing education	14,222	Cold	2000–2008
6	4	Clinical imaging and ultrasound	12,947	Stable	2009–2016
7	11	Medical professionalism	10,918	Cold	2000–2008
8	10	Interprofessional education	10,787	Stable	2009–2016
9	6	Pharmacy education	10,481	Stable	2009–2016
10	9	Burnout and wellbeing	10,401	Hot	2017–2024
11	8	Nutrition and public health education	10,129	Cold	2000–2008
12	7	Dental education	9,489	Cold	2000–2008
13	13	Surgical residency programs	8,702	Hot	2017–2024
14	15	Feedback and residency training	8,497	Stable	2009–2016
15	12	Diversity and equity	8,289	Hot	2017–2024

Notes: Trend classification based on Poisson GLM (offset for total annual output) with Newey–West autocorrelation-robust standard errors and FDR correction (primary specification; see Methods). Hot = significantly increasing relative proportion; Cold = significantly decreasing relative proportion; Stable = no significant change in relative proportion (absolute publication counts may still grow). Peak period = period with highest proportional share. [†]Topic 0 is reported for structural completeness only—it is a residual/background cluster whose trend classification and structural position are not substantively interpreted (see text); the same applies to the journal-metadata artifact clusters Topic 36 and Topic 38.

Table 2 Trend Classification of Research Topics (Poisson GLM with Offset, Newey–West HAC Standard Errors, BH-FDR Correction). IRR = Incidence Rate Ratio per Year (95% CI); Values >1 Indicate Increasing Proportional Share

Topic	Class	IRR (95% CI)	p (FDR)
16 (AI/ChatGPT)	Hot	1.088 (1.026–1.154)	0.010
34 (Climate/health)	Hot	1.072 (1.011–1.137)	0.035
29 (Empathy)	Hot	1.071 (1.058–1.084)	<0.001
25 (VR simulation)	Hot	1.068 (1.053–1.083)	<0.001
12 (Diversity/equity)	Hot	1.040 (1.021–1.060)	<0.001
33 (CAM/acupuncture)	Cold	0.915 (0.908–0.923)	<0.001
37 (TB/tuberculosis)	Cold	0.944 (0.934–0.953)	<0.001
18 (Psychiatry)	Cold	0.958 (0.953–0.964)	<0.001
8 (Public health ed)	Cold	0.960 (0.952–0.969)	<0.001
26 (Geriatrics)	Cold	0.964 (0.949–0.980)	<0.001
1 (Surgical skills)	Stable	0.996 (0.984–1.009)	0.690
6 (Pharmacy)	Stable	1.002 (0.994–1.010)	0.770
22 (Assessment)	Stable	1.000 (0.992–1.008)	0.990

Notes: Table 2 shows 13 representative rows—the top 5 Hot and top 5 Cold topics by IRR magnitude and 3 representative Stable topics—out of 38 substantive topics (10 Hot, 12 Cold, 16 Stable); Topic 0, the residual cluster, is also classified Cold but is not interpreted substantively (see Table 1). IRR = incidence rate ratio (change in a topic's relative share per year; an IRR of 1.05 means the topic's share of the field grows by about 5% annually). 95% CIs are Wald intervals from Newey–West HAC standard errors (maxlag = 3), which are robust to overdispersion and serial autocorrelation; p(FDR) = Benjamini–Hochberg-adjusted p-value. Full GLM results for all 39 topics, including the HC1 sensitivity specification, are in Supplementary Table S11.

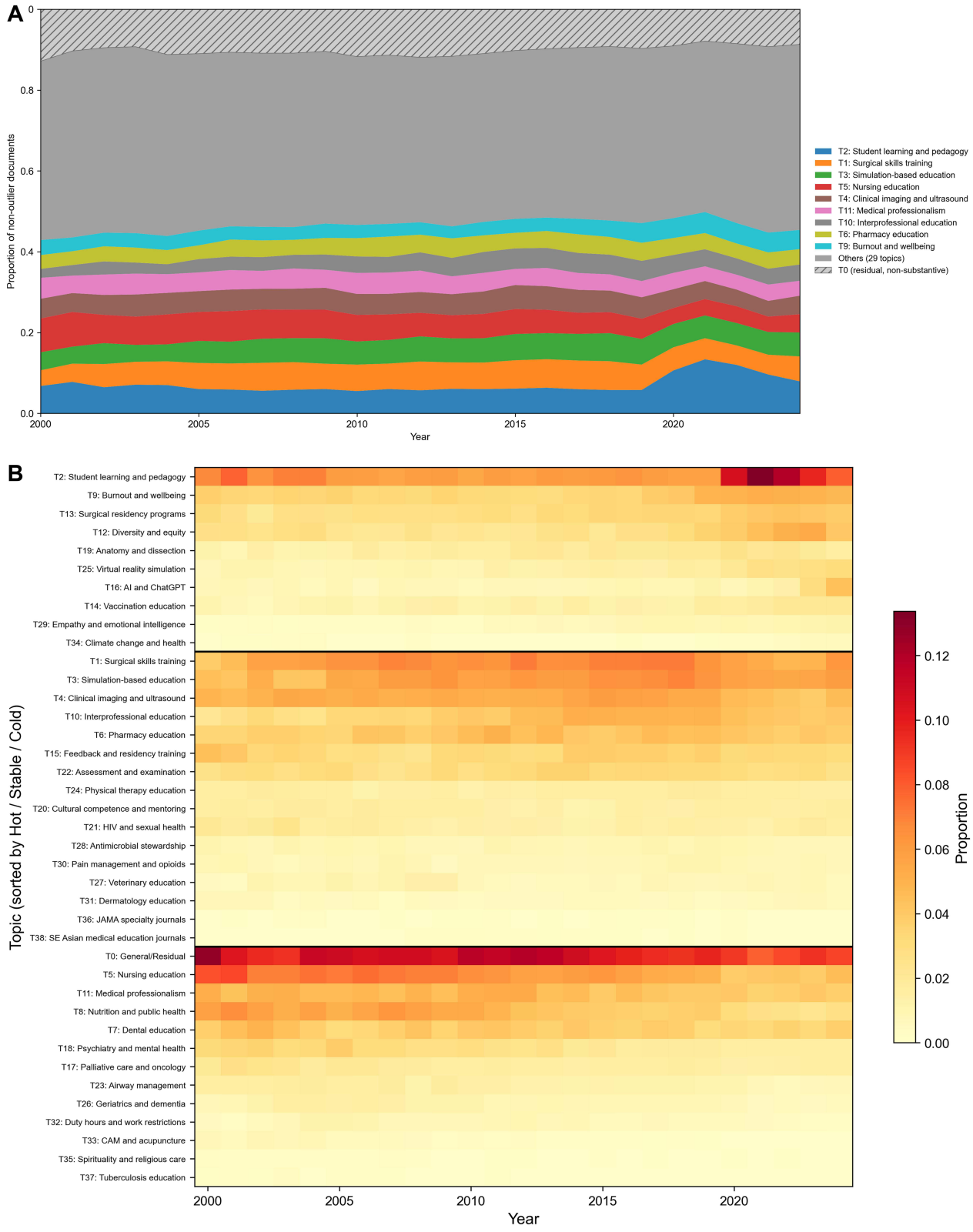

Figure 3 Temporal dynamics of topic proportions. (**A**) Stacked area chart showing the proportional share of the top-10 topics among non-outlier documents from 2000 to 2024. (**B**) Heatmap displaying the annual proportion of each of the 39 topics, grouped by the primary (Newey–West HAC) trend classification (Hot, Stable, Cold) and sorted by total document count within each group. Warmer colors indicate higher proportional representation in a given year; cell colors summarize cluster-level proportions, and individual clusters retain internal heterogeneity (see Limitations).

https://doi.org/10.2147/AMEP.S618140

Topic Stability and Robustness

Before presenting the structural positioning of topics, we summarize the robustness of the topic structure itself. Five-seed re-fitting of the BERTopic pipeline yielded a mean pairwise Adjusted Rand Index of 0.787 (SD = 0.048) on non-outlier documents, exceeding the prespecified stability threshold of 0.70. Varying the cosine-similarity threshold for outlier reassignment between 0.3 and 0.7 (primary value 0.5) preserved the substantive 39-topic structure across the entire range (Supplementary Table S13); the primary fit recorded 19,388 outlier documents (7.0% of the corpus) at threshold 0.5, with minor run-to-run variation around this value reflecting HDBSCAN edge-case stochasticity. Re-running the trend classification with Topic 0 excluded from both the model pool and the offset denominator changed none of the 38 remaining topic classifications (Supplementary Table S10), confirming that the Hot/Cold/Stable findings are not driven by the residual cluster; note also that in the primary analysis Topic 0 documents are part of the offset denominator, so its presence does not inflate the relative shares of substantive topics. Agreement between the primary HAC classification and the HC1 sensitivity specification was 35 of 38 substantive topics, with the three divergent topics (T3 Simulation, T4 Imaging, T10 Interprofessional) significant only under HC1; all topics with |beta| > 0.02 (eg., AI/ChatGPT, virtual reality simulation, burnout, diversity, climate, anatomy, vaccination) were classified identically under both specifications (Supplementary Table S4). An auxiliary sensitivity analysis of the exploratory interrupted time series that retained partial 2025 records is reported in Supplementary Table S3. Quadrant-threshold sensitivity for the embedding-based positional map (next subsection) showed 87.2% positional stability (34 of 39 topics retain their quadrant when the median split is replaced by a mean split; the 5 flipped topics are listed in Supplementary Table S14). Overall, these robustness checks indicate that the topic structure and the large-effect trend classifications are stable to reasonable variations in the analytic choices made along the BERTopic pipeline, while a small number of borderline trend and quadrant assignments are sensitivity-dependent and are flagged as such wherever they are discussed.

Structural Positioning Map

An embedding-based positional map characterizes topic positioning using time-invariant global centrality (mean cosine similarity between topic centroids) and density (mean pairwise cosine similarity among top-10 keyword embeddings); the quadrants are heuristic descriptors and not direct equivalents of the traditional Callon co-word framework (Figure 4 and Table 3). Established themes such as simulation (Topic 3), student learning (Topic 2), and nursing education (Topic 5) occupy the high-centrality high-density region, while rapidly evolving themes such as AI/ChatGPT (Topic 16) and burnout (Topic 9) sit in the high-centrality low-density region, consistent with topics that have not yet consolidated stable terminology. The complementary time-varying co-movement analysis (Supplementary Figure S3 and Supplementary Table S15) is reported as Supplementary Context only; we do not interpret quadrant migration substantively because of the threshold-dependence noted above.

Exploratory Temporal Reference Periods

As an auxiliary descriptive analysis, we examined whether topic-level publication rates exhibited temporal coincidence with major policy and technology reference periods (full results in Supplementary Table S16). Across the two temporally isolated reference points, WFME Global Standards (2003) and ACGME Milestones (2013), and the 2020–2022 compound period (COVID-19, USMLE Step 1 Pass/Fail, ChatGPT modeled jointly due to overlapping post-event windows), a high proportion of topics displayed concurrent level or slope shifts after FDR correction. This widespread significance most likely reflects shared long-term secular trends and unmodeled serial autocorrelation rather than event-specific effects, and we therefore use these analyses only to provide temporal context for the trend classification rather than as a basis for event attribution. The annual publication volume trajectory and reference periods are shown in Figure 1.

Discussion

This study applied BERTopic to 276,253 medical education publications spanning 2000–2024, identifying 39 research topics and revealing a gradual reorganization of the field's relative thematic emphasis alongside sustained growth in volume. The more than doubling of publication output from 2000–2008 to 2017–2024, combined with 10 Hot, 12 Cold,

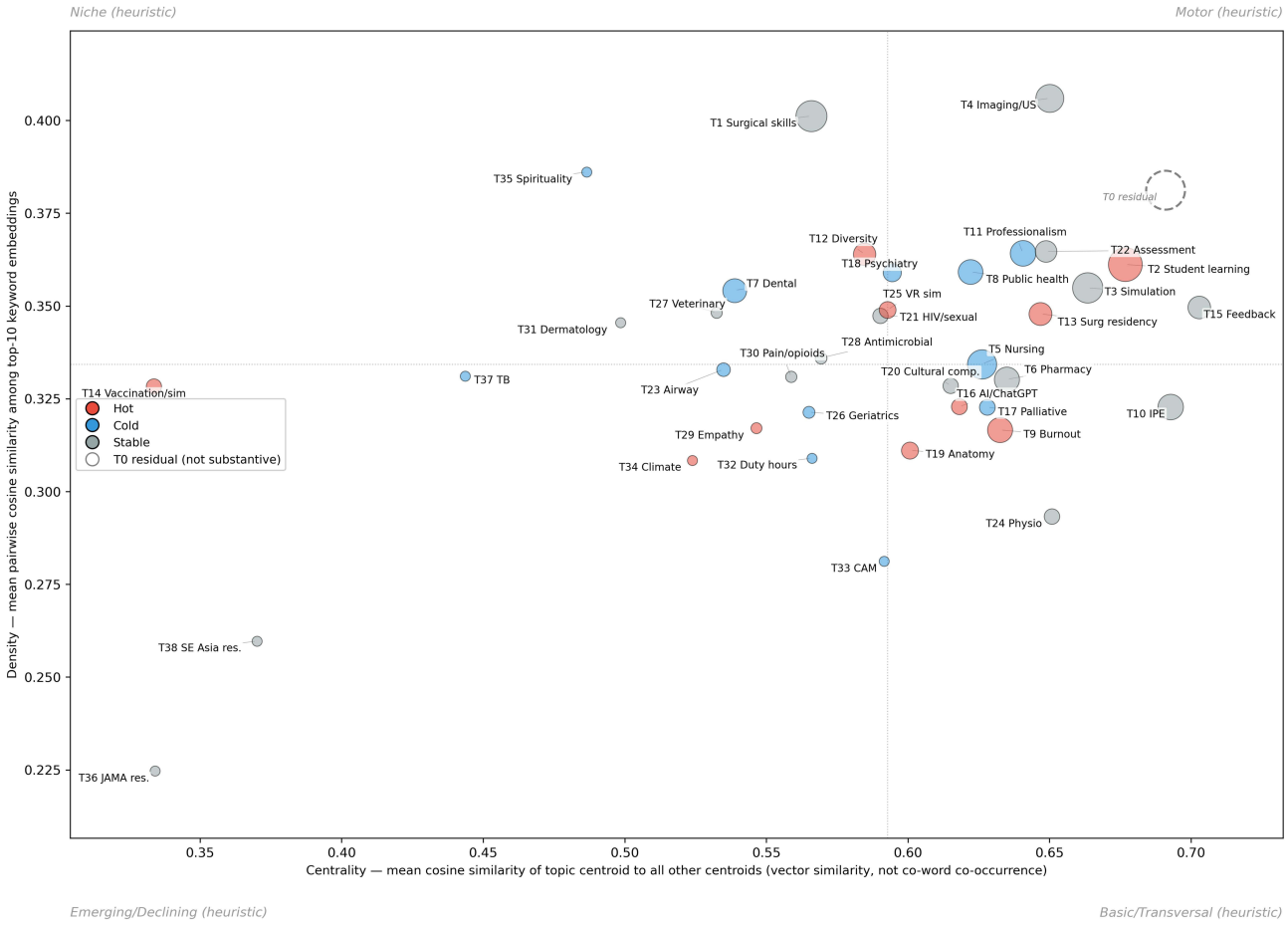

Figure 4 Embedding-based positional map. Each bubble represents one of 39 topics; bubble size is proportional to total document count. Both axes are time-invariant structural properties computed from the full-period embedding space: the x-axis (centrality) reflects mean cosine similarity of each topic's 384-dimensional centroid to all other topic centroids; the y-axis (density) reflects mean pairwise cosine similarity among each topic's top-10 keyword embeddings. Dashed lines indicate median centrality and density, dividing the space into four heuristic quadrants: Motor (high centrality, high density), Niche (low centrality, high density), Emerging/Declining (low centrality, low density), and Basic/Transversal (high centrality, low density). Quadrant labels are heuristic descriptors within the embedding-based framework and should not be equated with the traditional co-word network interpretation; assignments near the median thresholds are threshold-sensitive (Supplementary Table S14). Topics are colored by the primary (Newey–West HAC) trend classification (red = Hot, blue = Cold, gray = Stable). Topic 0 (Residual) is marked distinctly due to its limited interpretability. A complementary time-varying co-movement diagram is provided in Supplementary Figure S3.

and 16 Stable topics among the 38 substantive clusters under autocorrelation-robust inference, indicates that the field's research priorities are shifting at the margins, toward technology-enhanced and socially responsive themes, rather than undergoing wholesale paradigm replacement. We deliberately do not describe this shift as competency-based: competency-related content in our corpus is distributed across several clusters (surgical residency programs, feedback and residency training, assessment) rather than forming a distinct topic, so the topic-modeling results do not by themselves support claims about competency-based medical education as a unified research theme. This analysis extends previous

Table 3 Embedding-Based Structural Quadrant Assignments (Global, Time-Invariant)

Quadrant	N	Representative Topics
Motor (high C, high D)	12	Simulation (3), Student learning (2), Nursing (5), Surgical residency (13), VR (25)
Basic/Transversal (high C, low D)	8	Burnout (9), IPE (10), AI/ChatGPT (16), Palliative care (17), Anatomy (19)
Niche (low C, high D)	8	Surgical ed (1), Dental (7), Diversity (12), HIV/sexual health (21)
Emerging/Declining (low C, low D)	11	Vaccination (14), Empathy (29), Climate (34), Duty hours (32), CAM (33)

Notes: Centrality = mean cosine similarity of topic centroid to all other centroids (time-invariant); Density = mean pairwise cosine similarity of top-10 keyword embeddings (time-invariant). Quadrant assignment uses median centrality and density as thresholds. Full time-varying co-movement migration data in Supplementary Table S15.

bibliometric studies of medical education, which relied on keyword-based content analysis or citation rankings of substantially smaller corpora,[3,13,14] and the semantic embedding approach enabled finer-grained topic separation: for example, distinguishing virtual reality simulation (Topic 25) from general simulation-based education (Topic 3), and separating burnout/wellbeing (Topic 9) from broader professionalism research (Topic 11). Where Rotgans' content analysis of six core journals (1988–2010) identified student assessment, clinical and communication skills, clinical clerkships, and problem-based learning as its most prominent themes,[13] our full-corpus map confirms that these foundational themes remain structurally central while documenting the newer technology- and society-oriented growth areas that postdate that period.

Among the Hot topics, the technology cluster is the most prominent. Virtual reality simulation (Topic 25) showed one of the largest effect sizes, and AI and ChatGPT in education (Topic 16) emerged as a distinct, rapidly growing topic, evidence of how quickly research has responded to generative AI technologies.[22,23] General simulation-based education (Topic 3), by contrast, was Stable under the primary autocorrelation-robust specification and significant only under the less conservative HC1 errors: after two decades of expansion driven by patient safety imperatives, duty hour restrictions, and demonstrated effectiveness in procedural skill acquisition,[24–26] its share of the field appears to have plateaued at a high level, with growth now concentrated in the immersive-technology segment. Anatomy and dissection (Topic 19) also emerged as Hot, potentially reflecting renewed interest in anatomical education methods alongside virtual and augmented reality applications.

A second pattern concerns the concurrent growth of burnout and wellbeing (Topic 9), empathy (Topic 29), diversity and equity (Topic 12), and climate and health education (Topic 34). These topics are conceptually related: each concerns the human, social, or planetary context of medical training rather than its technical content. Their simultaneous Hot classification, together with the relative decline of traditionally framed professionalism research (Topic 11), is consistent with a broader turn toward socially responsive medical education, in which learner wellbeing, social justice, and planetary health are treated as core educational outcomes rather than peripheral concerns.[1,27,28] Topic modeling cannot establish why these themes co-emerged, but the pattern aligns with calls for socially accountable health professions education following the Lancet Commission report[1] and with documented high levels of learner distress.[28]

The Cold classification of several traditionally prominent topics warrants careful interpretation: the Hot/Cold/Stable labels reflect changes in relative proportional share within the growing corpus, not changes in absolute publication volume. Because total annual output more than quadrupled over the study period, a Stable or even Cold relative share is fully compatible with rising absolute output, and a declining share does not indicate abandonment of a topic. Topic 0's Cold classification likely reflects progressive differentiation of generically worded publications into more specialized topics over time rather than genuine decline; excluding it from the analysis altogether left all 38 substantive classifications unchanged. The proportional contraction of nursing education (Topic 5) and geriatrics education (Topic 26) may similarly reflect relative shifts in research emphasis or migration to dedicated journals outside the medical education search scope, rather than diminished activity in these fields.[29,30]

BERTopic's application to a corpus exceeding 276,000 documents demonstrates the scalability of transformer-based topic modeling for large-scale bibliometric analysis.[5] The semantic embedding approach captures contextual meaning beyond bag-of-words representations, which is particularly advantageous for abstract-length scientific texts. A parallel Latent Dirichlet Allocation model fitted on the same corpus (k = 39, identical preprocessing) produced a qualitatively different topic structure: token-level Jaccard overlap between matched topics remained low (0.14 to 0.25), and the six BERTopic themes examined mapped to only five distinct LDA topics, with simulation-based education (Topic 3) and virtual reality simulation (Topic 25) collapsing into a single LDA topic (Supplementary Figure S1). The hierarchical clustering of topic centroids (Supplementary Figure S4) shows the same pair as sibling but distinct branches within one macro-cluster, illustrating the finer semantic resolution of the embedding approach. This comparison is reported as descriptive context rather than as a claim of universal superiority, since the two model families produce non-commensurable quality indices. The embedding-based positional map provided a structural complement to the trend classification, but it is a heuristic positional adaptation, not a co-word strategic diagram in the Callon tradition:[21] centrality and density here reflect embedding-space similarity and keyword coherence rather than co-occurrence network topology, and the quadrant labels are heuristic descriptors whose stability under alternative thresholds is quantified in Supplementary Table S14.

The practical value of this map lies in field-level monitoring rather than prescription. For curriculum planners and educators, the rapid expansion of AI/ChatGPT and virtual reality research signals an accelerating evidence base to draw on when introducing AI literacy and immersive simulation into curricula, while the growth of climate and health education anticipates emerging expectations for planetary health content. For educational leaders and accreditation bodies, the sustained rise of burnout, wellbeing, empathy, and diversity research strengthens the case for structural attention to learner wellbeing and equity rather than individual-level remediation alone. For research funders, the map distinguishes densely saturated areas (eg., general simulation) from emerging niches with thin evidence bases (eg., climate-health education), which can inform prioritization. These implications should be read as hypothesis-generating signals from one data-driven lens on the literature, to be triangulated with systematic reviews and stakeholder priorities rather than acted on in isolation.

Several aspects of our analysis require cautious interpretation, and the topic-modeling output should be understood as a single data-driven lens on the literature rather than a definitive or exhaustive representation of the field. The ITS analysis is auxiliary and exploratory and cannot establish causal relationships:[31,32] the high proportion of significant associations likely reflects shared long-term temporal trends rather than event-specific effects,[33] two reference events (ACGME Milestones, USMLE Step 1) are primarily North American in scope, and the temporal proximity of COVID-19, USMLE Step 1, and ChatGPT precludes attribution to individual events. We therefore use the ITS results only as temporal context and report them in Supplementary Tables. The S-curve lifecycle analysis classified 82% of fitted topics as Mature with near-perfect fit statistics; this mainly reflects the mathematical property that cumulating any non-negative series yields smooth curves that logistic functions fit well, so high R-squared values should not be read as evidence of lifecycle maturity, and these fits are reported as Supplementary Context only.

Several limitations should be considered. First, the corpus was constructed from four biomedical and multidisciplinary databases (PubMed, Web of Science, Scopus, OpenAlex) but did not include education-specific databases such as ERIC. As a result, pedagogy-focused medical education research published primarily in education journals indexed by ERIC (particularly methodologically oriented work in instructional design, learning theory, and qualitative classroom research) is likely underrepresented in our corpus relative to its true share in the broader field. The resulting corpus is therefore best interpreted as a large English-language multi-database medical education corpus rather than an exhaustive mapping of the entire field, and our data-sharing approach is consistent with the Five Safes framework for research data access,[34] whereby analytical code and aggregated topic-count tables are available on request while raw licensed abstracts are not redistributed. Second, the English-language restriction excluded approximately 45% of initial records; our findings primarily reflect the English-language literature and may not capture regionally specific themes. Third, topic labels were assigned through an internal independent two-author review rather than an external blinded multi-rater expert panel; while we report agreement and the empirical basis (representative titles in Supplementary Table S5, label-title lexical overlap confidence in Supplementary Table S6), formal external multi-expert validation would be required before any individual label is treated as definitive. Specifically, Topic 14 ("Vaccination education") was retained as a Hot topic but its top representative titles in Supplementary Table S5 reveal a mixed cluster in which computational-simulation, phylogenetic-modeling, and machine-learning papers are interleaved with vaccination-education papers; readers should treat the keyword-prominent label as one component of a heterogeneous cluster rather than as a faithful summary of all underlying documents. Topics 36 (JAMA-family journal metadata) and 38 (Southeast Asian medical-education journal names) are journal-indexing artifacts retained in the GLM pool for completeness but not interpreted substantively. Fourth, the BERTopic pipeline involves stochastic components (UMAP, HDBSCAN), and although stability analysis confirmed acceptable reproducibility (ARI = 0.787), slightly different structures could emerge under different seeds. Fifth, the Poisson GLM residuals of many topics displayed positive first-order autocorrelation characteristic of annual publication time series with nonlinear curvature (eg., pandemic-related bumps); our primary inference therefore used Newey–West HAC standard errors (maxlag = 3), which are robust to both overdispersion and serial dependence. HAC estimators nevertheless have imperfect small-sample properties with only 25 annual observations, and the bandwidth choice (maxlag = 3) is itself a modeling decision; the less conservative HC1 specification, reported as a sensitivity analysis (Supplementary Table S4), classifies three additional borderline topics (T3, T4, T10) as significant, and we flag these as dependent on the choice of error structure. Sixth, the auxiliary ITS analysis tested 390 hypotheses across

overlapping event windows with limited annual data points, serial autocorrelation not explicitly modeled, and results should not be interpreted causally. Seventh, the embedding-based positional map uses embedding-space centrality and density rather than traditional co-word network metrics; future work could construct period-specific co-occurrence networks for more direct comparability with the Callon tradition. Finally, the all-MiniLM-L6-v2 embedding model was selected for computational efficiency and cross-disciplinary semantic balance rather than as a claim of superiority over domain-specific models such as PubMedBERT or BioBERT; comparative experiments using domain-specific embeddings would be a valuable extension of this work. The embedding model is also static: it encodes word and phrase meanings as learned at training time and cannot track semantic drift in educational terminology across the 25-year window (eg., evolving usage of "virtual patient" or "distance learning"), so the analysis captures the temporal distribution of documents in a fixed semantic space rather than the evolution of meaning itself.

Conclusions

This study demonstrates that large-scale transformer-based topic modeling, combined with autocorrelation-robust trend models and explicit sensitivity analyses, can map the thematic structure of a complete multi-database medical education corpus, a methodological contribution applicable to bibliometric research well beyond this field. Substantively, the relative emphasis of medical education research has shifted gradually toward technology-enhanced and socially responsive themes (artificial intelligence, virtual reality, wellbeing, empathy, diversity, and climate-health education), with 10 of 38 substantive topics accelerating and 12 contracting in relative share. Foundational topics such as assessment, surgical training, and simulation remain structurally central and continue to grow in absolute volume: increases in relative prominence of newer themes do not imply abandonment or decline of foundational educational topics. These classifications, and the heuristic positional map that accompanies them, depend on the modeling assumptions documented in our sensitivity analyses and should be read as one data-driven lens on the field rather than a definitive taxonomy. For educators, curriculum planners, accreditation bodies, and funders, the resulting map offers an empirical starting point for curriculum renewal, learner wellbeing and equity initiatives, and research prioritization. Future work should extend this approach with dynamic topic models that allow topic content to evolve over time, longitudinal validation against expert panels and qualitative analyses, comparisons with domain-specific embedding models, and corpora that incorporate education-specific databases.

Abbreviations

BERTopic, Bidirectional Encoder Representations from Transformers Topic modeling; UMAP, Uniform Manifold Approximation and Projection; HDBSCAN, Hierarchical Density-Based Spatial Clustering of Applications with Noise; c-TF-IDF, class-based Term Frequency-Inverse Document Frequency; LDA, Latent Dirichlet Allocation; ITS, Interrupted Time Series; GLM, Generalized Linear Model; FDR, False Discovery Rate; ARI, Adjusted Rand Index; NMI, Normalized Mutual Information; MMR, Maximal Marginal Relevance; ACGME, Accreditation Council for Graduate Medical Education; WFME, World Federation for Medical Education; AI, Artificial Intelligence; WoS, Web of Science.

Data Sharing Statement

The analytical code, BERTopic model parameters, and aggregated annual topic-count tables that do not contain copyright-restricted content are available from the corresponding author upon reasonable request. The raw bibliometric records cannot be redistributed directly due to database licensing restrictions, but full search queries for each database are provided in Supplementary Table S1b to enable independent replication.

Ethics Approval and Informed Consent

This study analyzed publicly available bibliometric records from PubMed, Web of Science, Scopus, and OpenAlex databases. No human subjects were involved, and no personally identifiable information was collected. Ethical approval was therefore not required for this study.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Conceptualization: C.C., T.Z. Data curation: C.C., J.L., Z.Z. Formal analysis: C.C., J.L., Z.Z. Investigation: J.L., Z.Z. Methodology: C.C. Project administration: T.Z. Software: C.C. Supervision: T.Z. Validation: J.L., Z.Z. Visualization: C.C., J.L. Writing – original draft: C.C., J.L., Z.Z. Writing – review & editing: C.C., J.L., Z.Z., T.Z. C.C., J.L., and Z.Z. contributed equally to this work as co-first authors. The study involved three major analytical components (BERTopic model training with topic-label review, Poisson GLM trend classification with interrupted time series auxiliary analysis, and S-curve lifecycle modeling with strategic positional mapping), each independently led by one co-first author. Specifically, C.C. designed the overall study framework, developed the BERTopic pipeline (UMAP, HDBSCAN, c-TF-IDF), implemented the Poisson GLM trend classification and embedding-based positional heuristic, created all main-text visualizations, and drafted the Introduction, Methods, and Discussion sections. J.L. independently conducted systematic literature searches across PubMed, Web of Science, Scopus, and OpenAlex, led the independent topic-label review (39 topics, two-stage protocol), contributed to figure design for the trend panels, and drafted the Results section. Z.Z. independently led the cross-database record merging across four databases (622,407 unique records from 943,925 raw retrievals) and tri-gram Jaccard deduplication, performed the S-curve lifecycle modeling and interrupted time series auxiliary analysis across policy and technology reference periods, and drafted the Supplementary Materials (15 tables, 4 figures). All three co-first authors participated in data curation, formal analysis, validation, and manuscript writing. T.Z. conceived the research direction, supervised the project, and critically revised the manuscript for intellectual content. All authors approved the final version.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Disclosure

The authors declare no competing interests related to this work.

References

1. Frenk J, Chen L, Bhutta ZA. et al. Health professionals for a new century: transforming education to strengthen health systems in an interdependent world. *Lancet*. 2010;376:1923–1958. doi:10.1016/S0140-6736(10)61854-5
2. Ninkov A, Frank JR, Maggio LA. Bibliometrics: methods for Studying Academic Publishing. *Perspect Med Educ*. 2022;11:173–176. doi:10.1007/s40037-021-00695-4
3. Lee K, Whelan JS, Tannery NH, Kanter SL, Peters AS. 50 years of publication in the field of medical education. *Med Teach*. 2013;35:591–598. doi:10.3109/0142159X.2013.786168
4. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *J Mach Learn Res*. 2003;3:993–1022.
5. Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure. *arXiv [Preprint]*. 2022. arXiv:2203.05794. doi:10.48550/arXiv.2203.05794
6. Grootendorst M. MaartenGr/BERTopic: v0.16. *Zenodo*. 2023. doi:10.5281/zenodo.10208607
7. Egger R, Yu J. A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts. *Front Sociol*. 2022;7:886498. doi:10.3389/fsoc.2022.886498
8. Abdelrazek A, Eid Y, Gawish E, Medhat W, Hassan A. Topic modeling algorithms and applications: a survey. *Inf Syst*. 2023;112:102131. doi:10.1016/j.is.2022.102131
9. Stammbach D, Zouhar V, Hoyle A, Sachan M, Ash E. Revisiting automated topic model evaluation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*; 2023; Singapore, ACL. 9348–9357. doi:10.18653/v1/2023.emnlp-main.581
10. Angelov D. Top2Vec: distributed Representations of Topics. *arXiv [Preprint]*. 2020. arXiv:2008.09470. doi:10.48550/arXiv.2008.09470
11. Ebadi A, Xi P, Tremblay S, Spencer B, Pall R, Wong A. Understanding the temporal evolution of COVID-19 research through machine learning and natural language processing. *Scientometrics*. 2021;126:725–739. doi:10.1007/s11192-020-03744-7
12. Thurzo A, Varga I. Revisiting the role of review articles in the age of AI-agents: integrating AI-reasoning and AI-synthesis reshaping the future of Scientific Publishing. *Bratisl Med J*. 2025;126:381–393. doi:10.1007/s44411-025-00106-8
13. Rotgans JI. The themes, institutions, and people of medical education research 1988–2010: content analysis of abstracts from six journals. *Adv Health Sci Educ Theory Pract*. 2012;17:515–527. doi:10.1007/s10459-011-9328-x
14. Azer SA. The top-cited articles in medical education: a bibliometric analysis. *Acad Med*. 2015;90:1147–1161. doi:10.1097/ACM.0000000000000780

15. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Hong Kong, China. ACL, 2019; pp. 3982–3992. doi:10.18653/v1/D19-1410
16. McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. *arXiv [Preprint]*. 2018. arXiv:1802.03426. doi:10.48550/arXiv.1802.03426
17. Campello RJGB, Moulavi D, Sander J. Density-based clustering based on hierarchical density estimates. In: Pei J, Tseng VS, Cao L, Motoda H, Xu G, editors. *Advances in Knowledge Discovery and Data Mining (PAKDD 2013). Lecture Notes in Computer Science*. Vol. 7819. Berlin/Heidelberg: Springer; 2013: 160–172. doi:10.1007/978-3-642-37456-2_14
18. Carbonell J, Goldstein J. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98); 1998; Melbourne, Australia, ACM, 335–336. doi:10.1145/290941.291025
19. World Federation for Medical Education. *Basic Medical Education: WFME Global Standards for Quality Improvement*. Copenhagen: WFME; 2003.
20. Nasca TJ, Philibert I, Brigham T, Flynn TC. The next GME accreditation system - rationale and benefits. *N Engl J Med*. 2012;366:1051–1056. doi:10.1056/NEJMsr1200117
21. Callon M, Courtial JP, Laville F. Co-word analysis as a tool for describing the network of interactions between basic and technological research: the case of polymer chemistry. *Scientometrics*. 1991;22:155–205. doi:10.1007/BF02019280
22. Wartman SA, Combs CD. Medical education must move from the information age to the age of artificial intelligence. *Acad Med*. 2018;93:1107–1109. doi:10.1097/ACM.0000000000002044
23. Masters K. Artificial intelligence in medical education. *Med Teach*. 2019;41:976–980. doi:10.1080/0142159X.2019.1595557
24. McGaghie WC, Issenberg SB, Petrusa ER, Scalese RJ. A critical review of simulation-based medical education research: 2003–2009. *Med Educ*. 2010;44:50–63. doi:10.1111/j.1365-2923.2009.03547.x
25. Cook DA, Hatala R, Brydges R, et al. Technology-enhanced simulation for health professions education: a systematic review and meta-analysis. *JAMA*. 2011;306:978–988. doi:10.1001/jama.2011.1234
26. Issenberg SB, McGaghie WC, Petrusa ER, Lee Gordon D, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. *Med Teach*. 2005;27:10–28. doi:10.1080/01421590500046924
27. Rotenstein LS, Torre M, Ramos MA, et al. Prevalence of burnout among physicians: a systematic review. *JAMA*. 2018;320:1131–1150. doi:10.1001/jama.2018.12777
28. Dyrbye LN, West CP, Satele D, et al. Burnout among U.S. medical students, residents, and early career physicians relative to the general U.S. *Population Acad Med*. 2014;89:443–451. doi:10.1097/ACM.0000000000000134
29. Bilimoria KY, Chung JW, Hedges LV, et al. National cluster-randomized trial of duty-hour flexibility in surgical training. *N Engl J Med*. 2016;374:713–727. doi:10.1056/NEJMoa1515724
30. Desai SV, Asch DA, Bellini LM, et al. Education outcomes in a duty-hour flexibility trial in internal medicine. *N Engl J Med*. 2018;378:1494–1508. doi:10.1056/NEJMoa1800965
31. Gordon M, Patricio M, Horne L, et al. Developments in medical education in response to the COVID-19 pandemic: a rapid BEME systematic review: BEME guide no. 63. *Med Teach*. 2020;42:1202–1215. doi:10.1080/0142159X.2020.1807484
32. Daniel M, Gordon M, Patricio M, et al. An update on developments in medical education in response to the COVID-19 pandemic: a BEME scoping review: BEME guide no. 64. *Med Teach*. 2021;43:253–271. doi:10.1080/0142159X.2020.1864310
33. Holmboe ES, Sherbino J, Englander R, Snell L, Frank JR; ICBME Collaborators. A call to action: the controversy of and rationale for competency-based medical education. *Med Teach*. 2017;39:574–581. doi:10.1080/0142159X.2017.1315067
34. Desai T, Ritchie F, Welpton R. Five safes: designing data access for research. In: *Economics Working Paper Series 1601*. Bristol: University of the West of England; 2016.