Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Mapping the Evolution of Artificial Intelligence in Healthcare, 2017–2025: An Integrated Bibliometric and Latent Dirichlet Allocation Topic Modelling Analysis

Piyusha Ambradkar

piyusha.a@somaiya.edu

K J Somaiya Institute of Management, Somaiya Vidyavihar University, Mumbai.

Vandana Tandon Khanna.

K J Somaiya Institute of Management, Somaiya Vidyavihar University, Mumbai.

Research Article

Keywords: artificial intelligence, healthcare informatics, bibliometric analysis, Latent Dirichlet Allocation, topic modelling, science mapping, federated learning, PRISMA 2020

Posted Date: September 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10263481/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

This study maps the intellectual evolution of artificial intelligence (AI) research in healthcare from 2017 to 2025 by integrating a standard seven-phase bibliometric pipeline with computational Latent Dirichlet Allocation (LDA) topic modelling, addressing the limitation that keyword co-occurrence analysis is constrained by author keyword heterogeneity. A dual-database corpus was extracted from OpenAlex (concept identifiers C154945302 and C71924100) and PubMed, screened under PRISMA 2020 to an AI-focused sub-corpus of 2,631 documents drawn from a 6,764-document parent cloud-healthcare dataset. Standard bibliometric methods (Bradford's Law, Lotka's Law, citation analysis, co-authorship and co-citation networks, bibliographic coupling, thematic mapping and evolution, Multiple Correspondence Analysis) were applied alongside LDA topic modelling (R topicmodels package, k = 8, Gibbs sampling, perplexity-optimised, Cohen's kappa = 0.84 for inter-rater topic labelling agreement). LDA identified eight coherent topics, two of which — large language models for clinical documentation and explainable AI governance — were found to be substantially under-represented in standard keyword co-occurrence clusters, demonstrating the added semantic discovery value of the integrated approach over keyword analysis alone. Topic trajectory analysis revealed a maturation gradient from deployment-ready applications (diagnostic imaging, Internet of Things-based clinical decision support) to investment-watch frontiers (large language models, explainable artificial intelligence) requiring governance development before clinical deployment. The findings provide a replicable integrated bibliometric-LDA methodology and a technology-readiness taxonomy directly applicable to health technology investment and horizon-scanning decisions.

Introduction

Artificial intelligence has undergone a transformation within healthcare that mirrors its broader technological trajectory: from narrow expert systems and rule-based clinical decision tools in the 1990s and 2000s, to machine learning models for predictive analytics in the 2010s, to the current generation of deep learning, federated learning, and large language model applications that are actively reshaping clinical workflows. The pace and scope of this transformation have made it genuinely difficult for researchers, technology managers, and health system decision-makers to maintain an accurate picture of the field's current state and near-term direction.

Bibliometric analysis provides a macro-level map of the scholarly landscape, identifying leading contributors, thematic clusters, and citation networks. Standard keyword co-occurrence analysis, however, is constrained by keyword heterogeneity: different authors may use different terms for the same concept, or identical terms for different concepts, introducing noise into the thematic signal. Latent Dirichlet Allocation (LDA) topic modelling addresses this limitation by discovering latent semantic themes directly from full-text content, specifically titles and abstracts, without relying on author-supplied keywords. The integration of LDA with a complete bibliometric pipeline therefore offers a methodological advance over either approach used alone.

This paper applies this integrated methodology to an artificial intelligence-in-healthcare sub-corpus derived from a parent dual-database cloud-healthcare dataset, focusing on the 2017–2025 period during which artificial intelligence applications in healthcare transitioned from proof-of-concept demonstrations to active clinical deployment. The remainder of the paper is organised as follows. Section 2 reviews prior bibliometric and topic-modelling approaches in this domain. Section 3 describes the integrated methodology in full. Sections 4 and 5 present bibliometric and LDA results respectively. Section 6 presents thematic evolution and emerging frontier analysis. Section 7 discusses implications for health technology management. Section 8 concludes with limitations and future research directions.

Literature Review

Bibliometric Approaches in AI Healthcare Research

The volume of bibliometric analyses of artificial intelligence in healthcare has grown substantially since 2019, driven by the availability of large-scale open bibliographic databases and reproducible R-based analytical pipelines [1]. Prior studies have documented the rapid growth of machine learning applications in radiology, oncology, and cardiology, and have identified citation networks linking artificial intelligence methodology development to clinical application domains. However, most prior bibliometric analyses of artificial intelligence in healthcare have relied on keyword co-occurrence as their primary thematic discovery method, an approach subject to keyword standardisation limitations, inconsistency across disciplinary communities, and an inability to capture semantic relationships between concepts described using different terminology in different research traditions.

Latent Dirichlet Allocation in Research Synthesis

LDA, introduced by Blei, Ng, and Jordan [2], models a corpus of documents as mixtures of latent topics, where each topic is represented as a probability distribution over vocabulary terms. Applied to scientific literature, LDA discovers coherent research themes from title and abstract text irrespective of the specific terminology individual authors employ. Prior applications of LDA to research synthesis include technology forecasting [3], patent landscape analysis, and systematic review augmentation. The integration of LDA with standard bibliometric pipeline methods, as proposed by Suominen and Toivanen [4], remains relatively uncommon in health informatics and had not previously been applied to the artificial intelligence-in-healthcare domain at the scale reported here.

Knowledge Gap

No prior study has applied an integrated bibliometric and LDA pipeline to an artificial intelligence-in-healthcare sub-corpus drawn from a dual-database OpenAlex and PubMed dataset spanning the 2017–2025 period. This gap is consequential because the most significant recent developments in artificial intelligence-enabled healthcare, including federated learning, large language models, and explainable

artificial intelligence, occurred predominantly after 2019; standard keyword analyses applied to older or narrower corpora are correspondingly likely to under-represent their semantic dimensions.

Methodology

Corpus Construction

The artificial intelligence-healthcare sub-corpus was derived from a parent 6,764-document cloud-healthcare dataset (extracted from OpenAlex, concept identifiers C2908647359 and C71924100, and PubMed, Title/Abstract fields and MeSH headings, 2010–2025) by filtering for documents containing artificial intelligence-related terms in title or abstract ("artificial intelligence", "machine learning", "deep learning", "neural network", "federated learning", "large language model", "natural language processing", "explainable AI") and restricting the result to publications dated 2017–2025. This produced a working sub-corpus of 2,847 documents. PRISMA 2020 screening, comprising deduplication via mergeDbSources() followed by year, language, and abstract-availability filters, yielded 2,631 eligible documents for full analysis. Table 1 reports the construction flow.

Table 1
PRISMA 2020 artificial intelligence-healthcare sub-corpus construction

Stage	Records (n)
Parent corpus (cloud and healthcare, 2010–2025)	6,764
AI-keyword filter applied (2017–2025)	2,847
After deduplication (mergeDbSources())	2,694
After language and abstract-availability filters	2,631
Final AI-healthcare analytical sub-corpus	2,631

Standard Bibliometric Pipeline

Standard bibliometric methods were applied to the 2,631-document sub-corpus using the bibliometrix package (version 4.x) in R [1]: Bradford's Law, Lotka's Law, global and local citation analysis, co-authorship network analysis with Louvain community detection (minimum two papers per author), co-citation network analysis, keyword co-occurrence analysis (minimum five co-occurrences), bibliographic coupling, thematic mapping via thematicMap(), thematic evolution via thematicEvolution() across three periods (2017–2019, 2020–2022, 2023–2025), and Multiple Correspondence Analysis via conceptualStructure() with a maximum of eight dimensions.

Latent Dirichlet Allocation Topic Modelling

LDA topic modelling was applied to the combined title and abstract text field using the topicmodels package in R [5]. Pre-processing comprised lowercase conversion, punctuation removal, English stop-word removal, a domain-specific stop-word list removing high-frequency non-discriminative terms (“patient”, “study”, “method”, “result”, “data”, “proposed”), and stemming via the SnowballC package. The Document-Term Matrix was constructed with a minimum term frequency of three. The optimal number of topics (k = 8) was selected by minimising held-out perplexity across a candidate range of k = 4 to k = 15. Gibbs sampling was used with 2,000 iterations and a burn-in of 500 iterations. Topics were labelled through qualitative review of the top twenty terms per topic by two independent coders, achieving an inter-rater agreement of Cohen's kappa = 0.84, indicating substantial agreement.

Use of Artificial Intelligence Tools in Manuscript Preparation

In accordance with the journal's policy on the disclosure of AI-tool use, the authors confirm the following. Latent Dirichlet Allocation, the topicmodels R package, and standard statistical computing constitute the analytical methodology under study and are not large language model writing assistance; their use is reported as the paper's substantive methodology rather than as a disclosed writing aid. No generative large language model was used to produce original analysis, findings, interpretation, or data. AI-assisted copy editing, as defined by the journal (wording and formatting improvements to author-generated text for grammar, spelling, and tone, without generative content creation), may have been used during manuscript preparation; the authors retain full accountability for the final text and confirm that all edits reflect their original analysis and conclusions.

Bibliometric Results

Production Trajectory

The 2,631-document sub-corpus exhibits explosive growth within the parent dataset's already exponential trend: annual artificial intelligence-healthcare publications grew from approximately 87 in 2017 to 731 in 2025, representing a compound annual growth rate of approximately 30 percent, slightly higher than the full parent dataset's approximate compound annual growth rate, consistent with artificial intelligence research having been a primary driver of the parent dataset's overall growth acceleration. The 2020 COVID-19 inflection is particularly pronounced in this sub-corpus: artificial intelligence-based COVID-19 diagnostics, particularly computed tomography image classification and federated surveillance models, generated a marked publication spike during 2020–2021.

Bradford's Law and Lotka's Law

Bradford's Law applied to the sub-corpus identifies IEEE Access, IEEE Internet of Things Journal, and Future Generation Computer Systems as Zone 1 core journals, consistent with the parent dataset and confirming that artificial intelligence and cloud-healthcare research share a common outlet ecosystem. Lotka's Law yields a beta coefficient of 2.58, marginally lower than the parent corpus's beta coefficient,

suggesting that the artificial intelligence sub-field has a slightly more concentrated core researcher community, consistent with the expectation that deep learning methodology expertise constitutes a higher barrier to entry than general cloud application research.

Citation Analysis

Table 2 presents the five most globally cited documents within the sub-corpus.

Table 2
Top five cited documents in the AI-healthcare sub-corpus

Rank	Title (abbreviated)	Journal	Year	Citations
1	FedHealth: a federated transfer learning framework for wearable health	IEEE Intelligent Systems	2020	954
2	Industry 4.0 and health: IoT, big data, and cloud computing for smart healthcare	Journal of Industrial Information Systems	2020	855
3	Smart healthcare: making medical care more intelligent	Global Health Journal	2019	694
4	HealthFog: deep learning smart healthcare for heart disease and stroke	Future Generation Computer Systems	2019	644
5	Deep-learning-enhanced human activity recognition for Internet of Health	IEEE Internet of Things Journal	2020	529

LDA Topic Modelling Results

Eight Discovered Topics

LDA topic modelling identified eight coherent latent topics within the sub-corpus, summarised in Table 3 alongside their corpus share and developmental phase assessment.

Table 3
Eight LDA topics: labels, top terms, corpus share, and developmental phase

Topic	Label	Top terms	% of corpus	Phase
T1	Deep learning: medical imaging	convolutional; image; classification; detection; neural; retinal	18.4%	Maturing
T2	NLP for clinical text	text; clinical; notes; extraction; EHR; named entity	13.2%	Growing
T3	Federated learning & privacy	federated; privacy; differential; aggregation; gradient; distributed	11.8%	Rapid growth
T4	IoT-integrated CDSS	IoT; sensor; wearable; fog; edge; monitoring; vital	16.7%	Maturing
T5	Drug discovery & genomics	genomic; drug; protein; sequence; molecular; omics	8.9%	Stable
T6	LLMs for clinical documentation	language model; summarisation; discharge; coding; prompt	9.3%	Explosive (2023+)
T7	Explainable AI governance	explainability; interpretability; SHAP; LIME; bias; fairness	10.6%	Fast growth
T8	AI remote patient monitoring	remote; wearable; chronic; readmission; telemonitoring	11.1%	Growing

Topic Trajectory Analysis

Topic proportion time series analysis, tracking each topic's proportional weight within the annual document set, reveals three distinct trajectory profiles. Topics T1 (deep learning imaging) and T4 (Internet of Things-based clinical decision support) show growth followed by a plateau, indicating field maturation in which the principal technical problems are substantially addressed and current publications focus on incremental improvement and clinical validation. Topics T2 (natural language processing for clinical text), T3 (federated learning), T7 (explainable artificial intelligence governance), and T8 (artificial intelligence-enabled remote monitoring) show sustained growth throughout the study period, indicating active frontier development. Topic T5 (drug discovery and genomics) shows stable, specialist-level volume. Topic T6 (large language models for clinical documentation) shows near-zero volume through 2022 followed by rapid acceleration during 2023–2025, a trajectory pattern characteristic of paradigm-shifting technology emergence, coinciding with the demonstrated clinical documentation competence of GPT-4 and subsequent large language models.

LDA-Bibliometric Integration

Cross-referencing LDA topics against keyword co-occurrence clusters reveals both alignment and informative divergence. LDA topic T3 (federated learning) aligns closely with the corresponding keyword co-occurrence cluster, providing mutual validation of both analytical methods. LDA topics T6 (large language models) and T7 (explainable artificial intelligence governance) are almost entirely absent from

keyword co-occurrence clusters, confirming that these represent genuine emerging semantic themes not yet captured by the established keyword vocabulary, and demonstrating the specific added value of LDA over keyword analysis for detecting nascent research fronts. LDA topic T2 (natural language processing for clinical text) is partially captured by the broader artificial intelligence and clinical decision support keyword cluster but is semantically richer, distinguishing between clinical natural language processing sub-domains, including entity extraction, summarisation, and coding, that collapse into a single keyword cluster under standard co-occurrence analysis.

Thematic Evolution and Emerging Frontiers

Thematic Map of the AI Sub-Corpus

The thematic map of the artificial intelligence-healthcare sub-corpus mirrors that of the parent dataset with greater granularity. Deep learning for medical imaging and Internet of Things-based clinical decision support occupy the motor-themes quadrant, indicating well-established, central research concerns. Electronic health record integration and cloud infrastructure constitute the basic-themes quadrant, representing the foundational layer underlying more specialised research. Drug discovery artificial intelligence and genomics occupy the niche-themes quadrant: high internal density but peripheral to the broader field. Federated learning, large language models, and explainable artificial intelligence governance occupy the emerging-themes quadrant, characterised by high growth velocity and increasing centrality.

Multiple Correspondence Analysis

Multiple Correspondence Analysis of the sub-corpus reveals three distinct conceptual dimensions. The first dimension, model architecture, aligns deep learning, convolutional neural network, transformer architecture, and model training terminology, representing the technical “how” of artificial intelligence systems. The second dimension, clinical application domain, groups imaging, natural language processing, remote monitoring, and clinical decision support terminology, representing the “where” of artificial intelligence deployment. The third dimension, governance and ethics, groups explainable artificial intelligence, bias, fairness, federated privacy, and regulatory compliance terminology, representing the dimension of “whether and how safely” artificial intelligence should be deployed, and which has grown most rapidly in relative prominence since 2020. The emergence of this third, governance-oriented dimension as a structuring force within the sub-corpus, clearly visible in the Multiple Correspondence Analysis but only partially visible in keyword co-occurrence analysis, constitutes the primary methodological contribution of the integrated LDA-bibliometric approach reported in this paper.

Discussion: Implications for Health Technology Management

An AI Readiness Taxonomy

The LDA topic trajectory analysis provides a technology management tool for assessing artificial intelligence application readiness for clinical deployment. Topic T1 (deep learning for medical imaging) is deployment-ready: the evidence base is mature, clinical validation exists across multiple disease areas including diabetic retinopathy screening, lung cancer detection, and electrocardiogram interpretation, and regulatory approvals are accumulating. Topic T4 (Internet of Things-based clinical decision support) is deployment-ready for sepsis prediction and chronic disease monitoring applications specifically. Topics T2, T3, and T8 are integration-ready: the underlying technology is proven, but deployment requires integration with existing electronic health record and monitoring infrastructure. Topics T6 and T7 are investment-watch frontiers: large language models for clinical documentation and explainable artificial intelligence governance are approaching deployment readiness but require regulatory framework development and institutional governance preparation before widescale clinical adoption.

Recommendations for Technology Managers and Researchers

Three recommendations follow from the analysis. First, federated learning, identified in this study as Topic T3, should be prioritised as the default architectural consideration for multi-site clinical artificial intelligence programmes, since it makes population-scale clinical models possible within existing data protection frameworks. Second, explainable artificial intelligence governance frameworks should be developed in advance of large-scale language model deployment; the explosive scholarly growth identified in Topic T7 typically precedes corresponding regulatory requirement development, and organisations that establish explainable artificial intelligence audit frameworks early will hold a governance advantage as deployment scales. Third, the LDA-bibliometric integration methodology demonstrated in this paper should be considered for adoption as a standard technology horizon-scanning approach by health technology management teams, since it identifies emerging semantic themes that manual keyword-based literature searches are demonstrated here to under-detect.

Conclusions, Limitations, and Future Research

Conclusions

This paper presents an integrated bibliometric and LDA topic modelling analysis of artificial intelligence in healthcare for the period 2017–2025, applied to a 2,631-document sub-corpus derived from a parent 6,764-document dual-database dataset. LDA identifies eight coherent topics, two of which, large language models for clinical documentation and explainable artificial intelligence governance, are found to be largely invisible to standard keyword co-occurrence analysis, demonstrating the methodological added value of the integrated approach over either method used alone. The topic trajectory analysis provides a technology management readiness taxonomy with direct application to health system artificial intelligence investment decisions. The study contributes a replicable integrated methodology and a current-state map of artificial intelligence in healthcare that warrants periodic updating as large language model and explainable artificial intelligence governance norms continue to consolidate.

Limitations

Four limitations should be noted. LDA results are sensitive to pre-processing choices and the selection of the number of topics; perplexity-based optimisation of the topic count is computationally principled but does not guarantee semantic coherence, and the topic labels reported here, although validated by substantial inter-rater agreement, remain an interpretive step subject to researcher judgement. The English-language restriction applied during corpus construction is particularly consequential for an artificial intelligence sub-corpus, since Chinese-language publications, predominantly appearing in journals not fully indexed in OpenAlex, represent a substantial share of global artificial intelligence-healthcare research output and are correspondingly under-represented in this analysis. Citation counts for the most recent publication years are necessarily incomplete given normal citation accumulation lags. Finally, this study characterises the structure and trajectory of the scholarly literature; it does not directly measure clinical deployment outcomes, and the technology-readiness taxonomy proposed in Section 7 should be understood as derived from publication-pattern evidence rather than from direct evaluation of deployed systems.

Future Research

Four directions for future research are proposed. Dynamic topic modelling could be applied to track fine-grained temporal evolution of the eight identified LDA topics with greater temporal resolution than the three-period thematic evolution analysis reported here. Comparison of LDA results against transformer-based topic modelling approaches, such as BERTopic, would help evaluate the semantic-depth trade-offs between classical probabilistic and embedding-based topic discovery methods. Citation network analysis explicitly linking LDA topics could trace cross-topic knowledge flows, for example identifying whether federated learning methodology development draws disproportionately on deep-learning-imaging citation networks. Finally, replication of the integrated bibliometric-LDA methodology on the drug discovery and genomics artificial intelligence sub-corpus, treated only briefly here as Topic T5, would extend this methodological contribution to a domain with distinct publication and citation dynamics.

Declarations

Conflicts of interest/Competing interests

The authors declare that they have no conflict of interest.

Ethics approval

This study is a bibliometric and computational text-analysis study of publicly available, de-identified scholarly metadata and did not involve human participants, human data, or animal subjects. Formal institutional ethics committee approval was therefore not required.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Use of artificial intelligence/large language models

Disclosed in full in the Methodology section above, in accordance with journal policy.

Funding

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

Author Contribution

Piyusha Ambradkar: conceptualisation, data curation, formal analysis, methodology, writing – original draft. Vandana Tandon Khanna: conceptualisation, supervision, writing – review and editing, validation.

Data Availability

The bibliometric corpus analysed in this study was extracted from OpenAlex (https://openalex.org) and PubMed (https://pubmed.ncbi.nlm.nih.gov), both publicly accessible databases. The screened analytical dataset, R analysis scripts, and the LDA topic-term matrices are available from the corresponding author upon reasonable request.

Code availability

Analysis was conducted using the publicly available R packages bibliometrix (version 4.x), topicmodels, and SnowballC. Custom analysis scripts are available from the corresponding author upon reasonable request.

References

1. Aria, M., Cuccurullo, C.: bibliometrix: an R-tool for comprehensive science mapping analysis. J. Informetr. **11**(4), 959–975 (2017)

2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
3. Yoon, J., Kim, J., Lee, C., et al.: Technology opportunity discovery using deep learning-based text mining and a knowledge graph. Technol. Forecast. Soc. Change. **180**, 121718 (2022)
4. Suominen, A., Toivanen, H.: Map of science with topic modeling: comparison of unsupervised learning and human-assigned subject classification. J. Assoc. Inf. Sci. Technol. **67**(10), 2464–2476 (2016)
5. Grün, B., Hornik, K.: topicmodels: an R package for fitting topic models. J. Stat. Softw. **40**(13), 1–30 (2011)
6. Page, M.J., McKenzie, J.E., Bossuyt, P.M., et al.: The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. **372**, n71 (2021)
7. Priem, J., Piwowar, H., Orr, R.: OpenAlex: a fully-open index of the world's research. arXiv ; (2022). 2205.01833.
8. Rahmani, A.M., Gia, T.N., Negash, B., et al.: Exploiting smart e-health gateways at the edge of healthcare Internet-of-Things: a fog computing approach. Future Gener Comput. Syst. **78**, 641–658 (2017)
9. van Eck, N.J., Waltman, L.: Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics. **84**(2), 523–538 (2010)
10. Zhu, H., Xiong, H., Zhang, Y., Wang, S., Zhou, Y.: FedHealth: a federated transfer learning framework for wearable healthcare. IEEE Intell. Syst. **36**(4), 83–93 (2020)