Commentary

Mapping the Scientific Agenda of Italian Oncology Nursing: A Topic Modelling Commentary on AIIAO 2026 Congress Abstracts

Citation: Caruso, R., De Maria, M., Conte, G., Arcidiacono, M. A., Caponetti, S., Cremona, G., Dabbene, M., Guberti, M., Piredda, A., & Magon, A. "Mapping the Scientific Agenda of Italian Oncology Nursing: A Topic Modelling Commentary on AIIAO 2026 Congress Abstracts" (2026) *International Journal of Health Supplements* 1 (Suppl. 2) 116-125. DOI: https://doi.org/10.82051/ijhs-2026-545

Published: July 7, 2026

Correspondence: Rosario Caruso – Associazione Italiana Infermieri di Area Oncologica (AIIAO), Italy; Mail: rosario.caruso@unimi.it

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Conflict of Interest: Rosario Caruso serves as President of the Italian Association of Oncology Nurses (AIIAO). Gianluca Conte, Marco Alfredo Arcidiacono, Serena Caponetti, Gabriele Cremona, Marcella Dabbene, Monica Guberti, Alessio Piredda, and Arianna Magon serve as members of the AIIAO Executive Board. Maddalena De Maria served as a member of the Scientific Committee of the AIIAO 2026 Congress. These institutional roles are disclosed as potential non-financial conflicts of interest. The authors declare no financial conflicts of interest or other relevant competing interests.

Funding: This research received no external funding.

Rosario Caruso[1,2], Maddalena De Maria[3], Gianluca Conte[1], Marco Alfredo Arcidiacono[4], Serena Caponetti[5], Gabriele Cremona[6], Marcella Dabbene[7], Monica Guberti[8], Alessio Piredda[9], Arianna Magon[10]

[1] Health Professions Research and Evidence Transfer Unit, IRCCS MultiMedica, Sesto San Giovanni, Italy
[2] Department of Biomedical Sciences for Health, Università degli Studi di Milano, Milan, Italy
[3] Department of Life, Health Sciences and Health Professions, Link Campus University, Rome, Italy
[4] Medical Department, University Hospital of Parma, Parma, Italy
[5] UOC Oncologia, Ospedale "A. Murri", AST, Fermo, Italia
[6] Department of Hematology-Oncology, Piacenza Hospital, Piacenza, Italy
[7] Nursing Management, La Maddalena S.p.A., Palermo, Italy
[8] Nursing and Allied Health Professions, IRCCS Istituto Ortopedico Rizzoli, Bologna, Italy
[9] Outpatient Service, IEO – European Institute of Oncology IRCCS, Milan, Italy
[10] Department of Public Health, Experimental and Forensic Medicine, Section of Hygiene, University of Pavia, Pavia, Italy

Abstract

Background. Scientific congresses act as observatories of a professional community's priorities and methods. The abstracts submitted to a national oncology nursing congress form a real-world corpus through which research and practice trajectories can be examined empirically rather than impressionistically.
Objective. To identify and interpret the emerging topics within the abstracts included in the AIIAO 2026 Congress proceedings using lexicometric analysis followed by topic modelling, and to discuss their implications for the oncology nursing agenda.
Methods. Exploratory, corpus-based text-mining study. The English structured bodies of the 82 included abstracts were preprocessed (lowercasing; removal of punctuation, digits and

structural headings; lemmatization; moderate stopword control; n-gram detection). Lexicometric descriptives were computed, and Latent Dirichlet Allocation models were estimated for k = 4-7. The number of topics was selected by combining coherence with the CaoJuan2009, Arun2010 and Deveaud2014 criteria and interpretability, not by automatic optimisation alone.

Results. The corpus contained 21,347 running words and 3,568 unique word forms (type-token ratio 0.167; 44.6% hapax). A five-topic solution was selected: care pathways, procedural safety and quality improvement (35.4%); self-management, treatment adherence and nursing surveillance (14.6%); nursing research capacity and symptom-distress measurement (19.5%); professional roles, leadership and organizational/ethical development (17.1%); and palliative and supportive care, quality of life and continuity (13.4%). The latent topics cut across the author-assigned editorial clusters.

Conclusions. The topics suggest a maturing research community oriented towards safe, standardized and person-centred care, supported by advanced roles, outcome measurement and digital health. The findings describe the submitted corpus and are not representative of Italian oncology nursing as a whole.

Keyword: Oncology Nursing, Topic Modelling, Care Pathways, Nursing Leadership, Palliative Care, Digital Health, Nursing Research

Introduction

Cancer remains one of the largest contributors to the global burden of disease, and the demand for skilled nursing care across the cancer trajectory continues to grow (Kocarnik et al., 2022). Within this context, oncology nursing has evolved from a predominantly task-oriented role into a research-engaged specialty with its own scientific priorities, advanced clinical roles, and policy ambitions (Rosenzweig et al., 2024). National congresses are a privileged vantage point on this evolution: the abstracts a professional community submits to its annual meeting represent, in aggregate, a snapshot of what its members are studying, implementing and prioritising at a given moment.

Reading such a corpus one abstract at a time is informative but inevitably selective, and narrative summaries of congress content tend to reflect the interests of whoever writes them. Automated text analysis offers a complementary, reproducible lens. Lexicometric analysis quantifies the vocabulary and phraseology of a corpus, while probabilistic topic modelling, most commonly Latent Dirichlet Allocation (LDA) (Blei et al., 2003), represents documents as mixtures of latent topics, each defined by a characteristic distribution of words. These techniques are increasingly applied to nursing texts, from clinical care reports (Danler et al., 2024) to the analysis of global research trends (Oh et al., 2024), precisely because they surface thematic structure that is difficult to perceive by manual reading alone.

The Italian Association of Oncology Nurses (AIIAO) 2026 Congress generated a substantial set of submitted abstracts spanning clinical innovation, specialist competencies, continuity of care, patient-reported outcomes, therapeutic education, digital health, and service organization. This corpus is an opportunity to characterise, with explicit and auditable methods, the themes that currently animate Italian oncology nursing. We treat the corpus as empirical material to map emerging topics, while remaining cautious about inferring the real-world prevalence of these phenomena in clinical practice.

Accordingly, the objective of this commentary is to identify and interpret the emerging topics within the abstracts included in the AIIAO 2026 Congress proceedings using lexicometric analysis followed by topic modelling analytics, and to

discuss their implications for the scientific agenda, specialist competencies, digital health, patient-reported outcomes and measures of experience, continuity of care, and the educational and policy priorities of oncology nursing and of AIIAO.

Methods

Study design

This is an exploratory, corpus-based text-mining (data-mining) study of congress abstracts. It is neither a systematic review nor a survey, and it makes no claim to statistical representativeness of Italian oncology nursing. The unit of analysis is the final abstract included in the AIIAO 2026 proceedings.

Corpus and unit of analysis

The corpus was derived from the AIIAO 2026 proceedings and the associated REDCap submission export. Of 91 submitted records inspected, 85 were complete; four incomplete records were excluded as not included in the final abstract book, and five duplicate records were removed after pre-specified duplicate-group checks (groups 26/27, 35/36, 53/68, and 75/89/90), yielding 82 unique abstracts (identifiers A001–A082). Each abstract appears once, irrespective of whether it was present in both the manuscript and the export. The analysed text was the English structured abstract body, which was available for all 82 abstracts; titles and keywords were not included in the primary model to avoid the noise introduced by mixed Italian–English titles. The corpus was therefore linguistically homogeneous (English).

Preprocessing

Each abstract body was lowercased; apostrophes, punctuation, and isolated digits were removed; and recurrent structural headings (introduction, background, methods, results, conclusions, discussion, objective, aim, purpose, keywords) were deleted as non-semantic noise. Tokens were lemmatized (WordNet), and English stopwords were removed. A moderate domain-specific control was applied for the topic model only: a small set of hyper-frequent, weakly discriminating terms (study, patient, oncology, nursing/nurse) was removed from the topic-modelling input while retained in the raw frequency counts, so that conceptually central terms were preserved in the descriptive lexicometry. Statistically detected bigrams and trigrams (for example, palliative care, quality of life, case manager) were retained. All preprocessing decisions were logged in an audit trail.

Lexicometric analysis

Descriptive lexicometry quantified the size and richness of the corpus: the number of abstracts, total running words (tokens), unique word forms (types), hapax legomena (words occurring once), the type–token ratio (TTR), and abstract-length statistics. We then examined the most frequent content terms, the most frequent bigrams and trigrams, and discriminating terms identified through term frequency–inverse document frequency (TF-IDF). Together, these indicators describe vocabulary volume and diversity and provide a foundation for topic modelling by confirming that the corpus is sufficiently varied to reveal meaningful themes.

Topic modelling and selection of the number of topics

Latent Dirichlet Allocation (LDA) (Blei et al., 2003) was used as an interpretable generative model consistent with a commentary. A document–term matrix was constructed from the preprocessed corpus, and models were estimated for k = 4, 5, 6 and 7 topics. The number of topics was not chosen by automatic optimisation alone. We combined topic coherence (Röder et al., 2015) with three established model-selection criteria—the density-based CaoJuan2009 criterion (minimised) (Cao et al., 2009), the Arun2010 symmetric Kullback–Leibler criterion (minimised) (Arun et al., 2010), and the Deveaud2014 information-divergence criterion (maximised)—together with the semantic interpretability of the resulting topics (Deveaud et al., 2014). The Griffiths2004 marginal-likelihood criterion is referenced as part of this family of metrics but was not computed in the present implementation (Griffiths & Steyvers, 2004); coherence was used as the interpretability-oriented indicator in its place, and this choice is stated transparently. Candidate solutions were presented with their metrics and top words before a final solution was agreed.

Interpretive validation and topic labelling

For the selected solution, each topic was characterised by its top words and bigrams, its prevalence (the proportion of abstracts assigned to it as the dominant topic), and its most representative abstracts (those with the highest topic probability). Interpretive labels were assigned by reading these representative abstracts alongside the top terms; the labels are interpretations of latent dimensions that may overlap and are not fixed categories. Each abstract was assigned to its dominant topic using document–topic probabilities, and a full assignment table (abstract identifier, title, dominant topic, maximum probability, and second topic) was retained as a supplementary audit output.

Software and reproducibility

Analyses were performed in Python version 3.12 for Windows using scikit-learn (TF-IDF), gensim (LDA, phrase detection, and coherence), and NLTK (lemmatization and stopwords). A fixed random seed was used for all models. The clean corpus, the analysis script, the lexicometric and topic-model outputs, the figures, and a process log are provided as supplementary materials to allow the analysis to be reproduced and audited.

Results

Corpus and lexicometric descriptives

The corpus comprised 82 abstracts spanning the nine author-assigned editorial clusters of the proceedings. It contained 21,347 running words and 3,568 unique word forms, with a TTR of 0.167 and 1,591 hapax legomena (44.6% of all word forms), indicating a varied, specialised vocabulary. Abstracts had a mean length of 260 tokens (median 248; range 158–1,432). The most frequent content terms were care, patient, nurse, clinical, and study; the most frequent bigrams were palliative care, quality of life, care pathway, continuity of care, and case manager.

TF-IDF highlighted discriminating terms, including symptom, distress, pathway, leadership, organizational, palliative, and research. Table 1 summarises these descriptives, and Figure 2 (panels 2a and 2b) shows the most frequent terms and bigrams.

Table 1. Corpus and lexicometric descriptives.

Indicator	Value
Abstracts analysed	82
Editorial clusters represented	9
Total tokens (running words)	21,347
Unique word forms (types)	3,568
Hapax legomena, n (%)	1,591 (44.6%)
Type–token ratio	0.167
Mean abstract length (tokens)	260.3
Median abstract length (tokens)	248
Length range (tokens)	158–1,432
Most frequent terms	care; patient; nurse; clinical; study
Most frequent bigrams	palliative care; quality of life; care pathway; continuity of care; case manager

Note. Lexicometric indicators are computed on cleaned surface word forms; the most frequent terms and bigrams are computed on lemmatized content tokens.

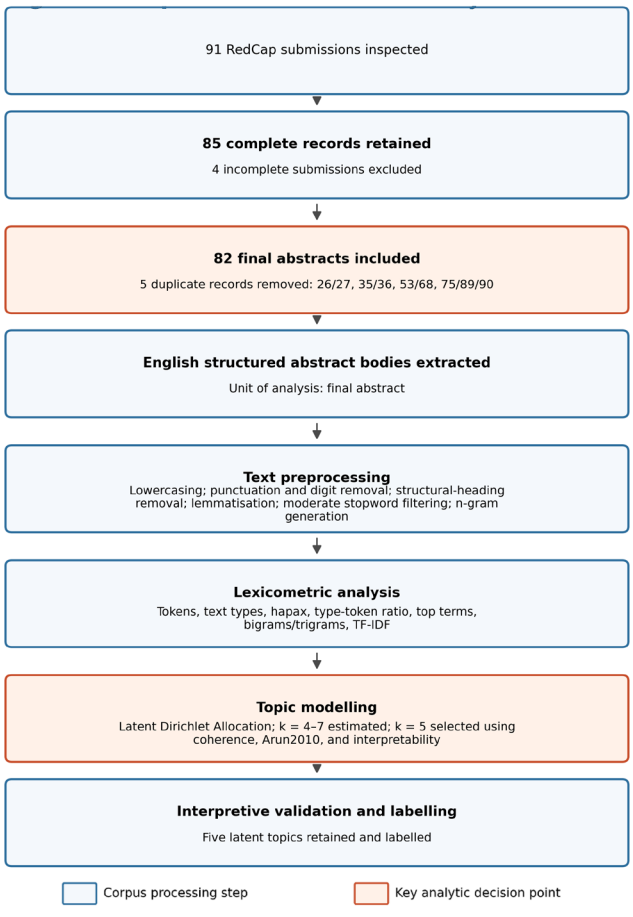

Figure 1. Corpus construction and analytic workflow

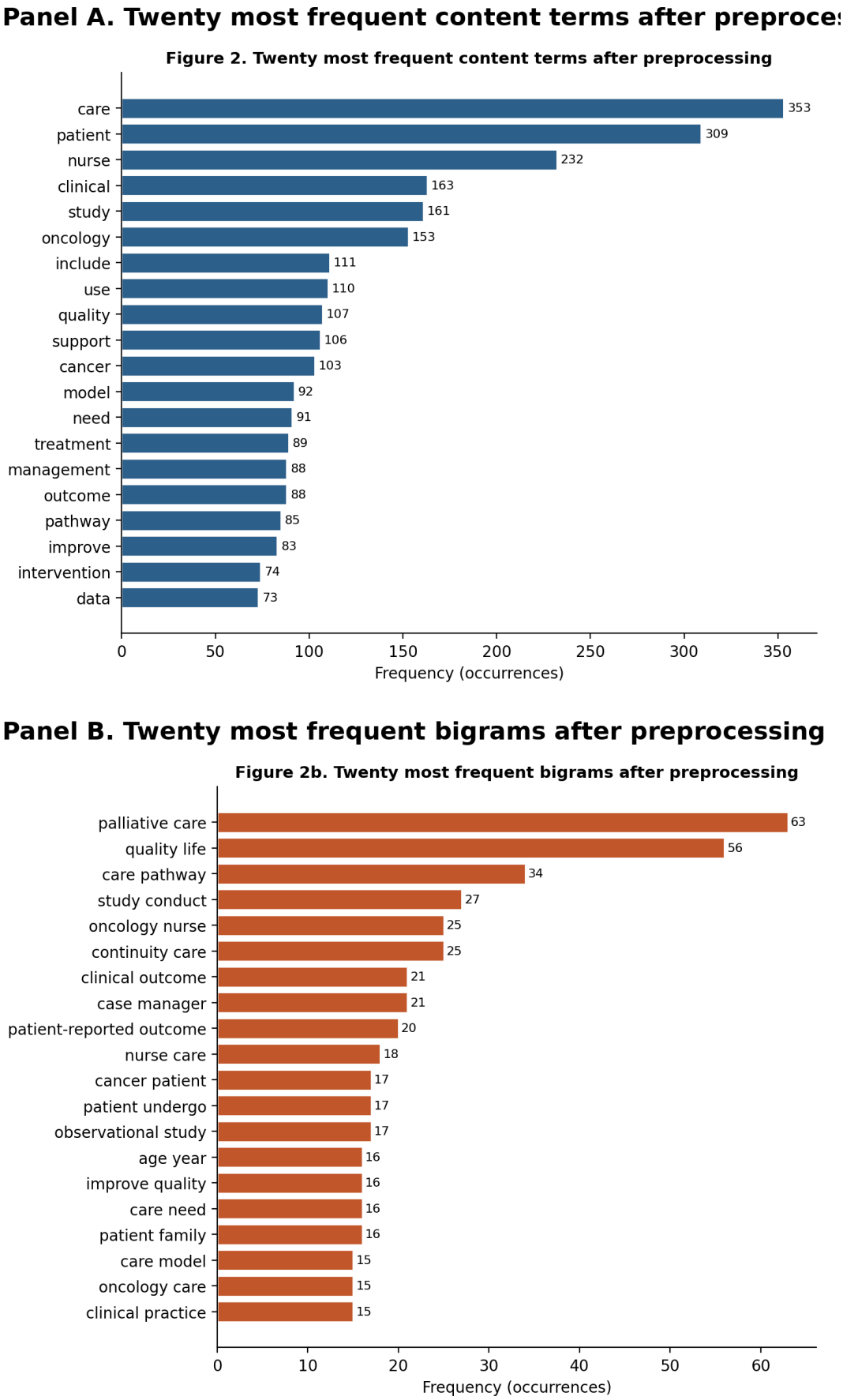

Figure 2. Lexicometric profile of the AIIAO 2026 abstract.

Note. Panel A shows the twenty most frequent content terms, while Panel B shows the twenty most frequent bigrams. Frequencies refer to occurrences in the preprocessed corpus and highlight the centrality of care, patients, nursing, oncology, quality of life, palliative care, care pathways, continuity of care, and patient-reported outcomes within the submitted abstracts.

Selection of the number of topics

Across the candidate solutions, the coherence-based and Arun2010 criteria favoured a five-topic model, whereas the CaoJuan2009 and Deveaud2014 criteria improved monotonically with more topics—a known tendency of these criteria to reward finer partitions. The seven-topic solution produced a very small topic (five abstracts) and lower coherence. Weighing the metrics against interpretability, a five-topic solution was selected as the main model (Table 2 reports the comparison).

Table 2. Corpus and lexicometric descriptives.

k	Coherence c_v (↑)	CaoJuan2009 (↓)	Arun2010 (↓)	Deveaud2014 (↑)
4	0.341	0.617	0.068	0.169
5	0.347	0.565	0.034	0.198
6	0.304	0.555	0.045	0.207
7	0.298	0.523	0.039	0.230

Note. The selected five-topic solution is highlighted. Arrows indicate the optimisation direction of each criterion.

Five-topic solution: labels, prevalence and top words

The five topics and their prevalence were: Topic 1, care pathways, procedural safety and quality improvement (29 abstracts, 35.4%); Topic 2, self-management, treatment adherence and nursing surveillance (12, 14.6%); Topic 3, nursing research capacity and symptom–distress measurement (16, 19.5%); Topic 4, professional roles, leadership and organizational/ethical development (14, 17.1%); and Topic 5, palliative and supportive care, quality of life and continuity (11, 13.4%). Figure 3 shows topic prevalence; Table 3 reports labels, top terms, representative abstracts, and interpretive meaning.

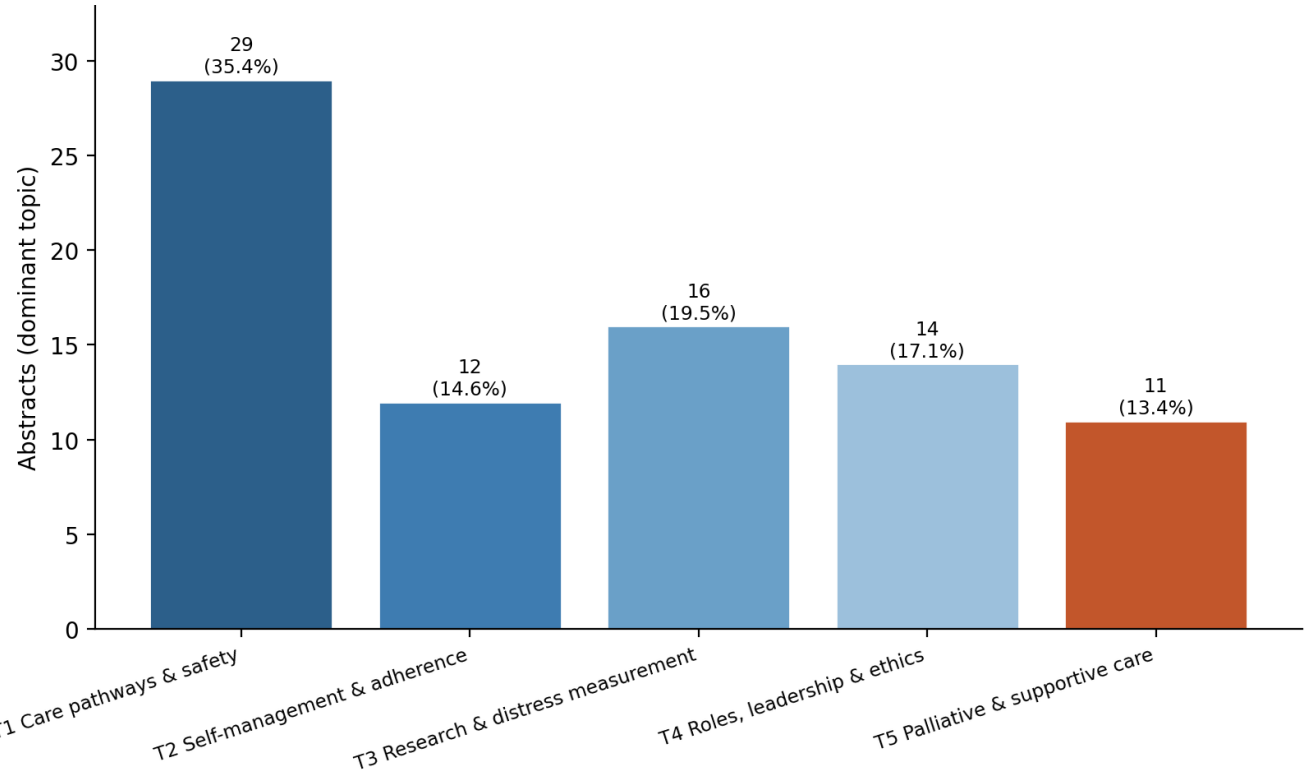

Figure 3. Topic prevalence (dominant-topic assignment, k = 5).

Table 3. Topic interpretation (five-topic solution).

Topic (label)	Prevalence n (%)	Top terms	Representative abstracts and interpretive meaning
T1. Care pathways, procedural safety and quality improvement	29 (35.4)	safety, improve, process, pathway, procedure, implementation, structure, tool, model	A006, A007, A019. Standardisation and safety of clinical processes, care-pathway design and quality-improvement initiatives.
T2. Self-management, treatment adherence and nursing surveillance	12 (14.6)	management, surveillance, self-care, monitor, adherence, perceive, intervention	A001, A002, A012. Adherence to (oral) anticancer therapy, self-management support and ongoing nursing monitoring.
T3. Nursing research capacity and symptom–distress measurement	16 (19.5)	research, symptom, distress, experience, network, assessment, protocol, questionnaire	A004, A008, A015. Research infrastructure and roles alongside instrument-based measurement of symptoms and distress.
T4. Professional roles, leadership and organizational/ethical development	14 (17.1)	professional, leadership, ethical, role, value, organizational, knowledge, practice	A024, A042, A049. Advanced roles, leadership, educational needs and organizational/ethical dimensions of practice.
T5. Palliative and supportive care, quality of life and continuity	11 (13.4)	treatment, palliative care, disease, pain, family, quality of life, score	A014, A022, A034. Palliative and supportive care, symptom and quality-of-life outcomes, and continuity across settings.

Note. Topics are latent dimensions and may overlap; representative abstracts are those with the highest topic probability.

Latent topics versus editorial clusters

The latent topics did not simply reproduce the nine author-assigned editorial clusters. As Figure 4 shows, each topic drew abstracts from several clusters: for example, Topic 1 aggregated abstracts from all clusters, while Topic 4 concentrated abstracts originally assigned to organization/policy and to the Bologna congress stream. This divergence is itself informative—it indicates that the vocabulary authors actually use cuts across the thematic labels they select at submission, and it should be read as a feature of the analysis rather than an inconsistency.

Discussion

Applying lexicometric analysis and topic modelling to the 82 abstracts of the AIIAO

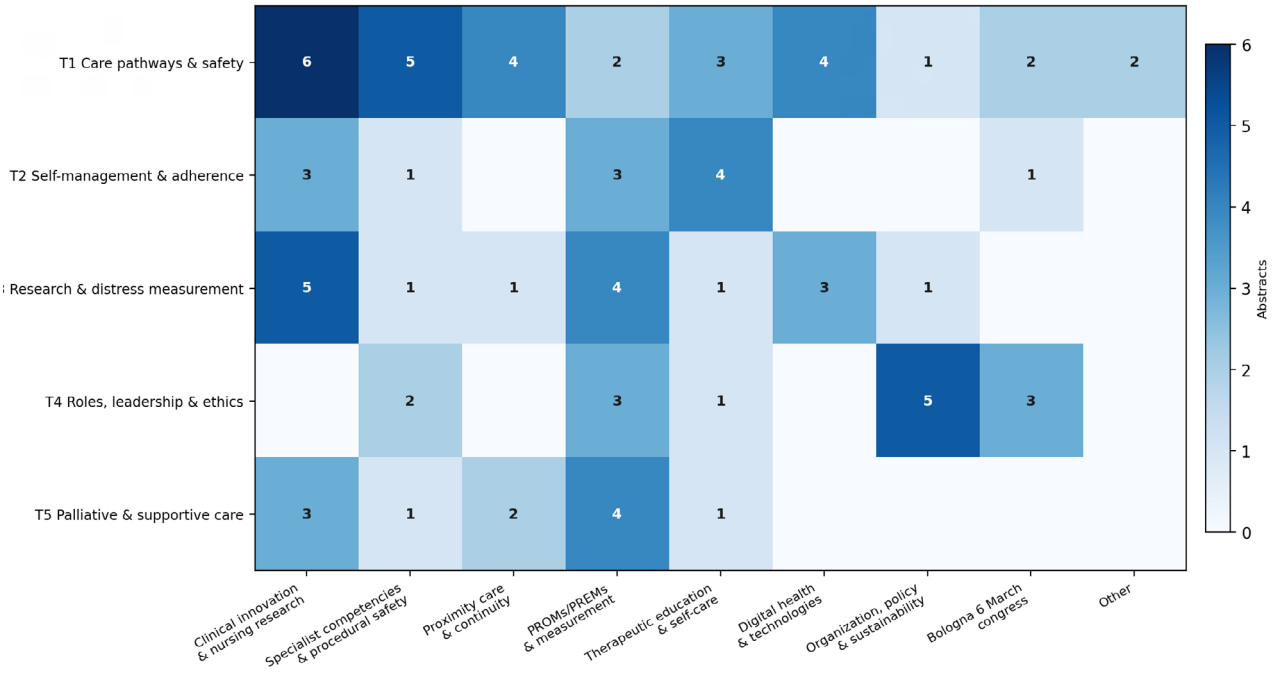

Figure 4. Latent topics (rows) versus author-assigned editorial clusters (columns).

2026 proceedings produced five interpretable topics: care pathways and procedural safety; self-management and adherence; research capacity and symptom–distress measurement; professional roles and leadership; and palliative and supportive care. The most prevalent topic concerned the design and safety of standardized care processes, and the recurrent bigrams, i.e., palliative care, quality of life, care pathway, continuity of care, case manager, reinforced a picture of a specialty preoccupied with organizing safe, coordinated, and person-centered cancer care.

Read as signals rather than measurements, these topics are consistent with a maturing research community. The prominence of research-capacity language and of instrument-based assessment (questionnaires, distress, symptom measurement) within Topic 3 suggests that methodological awareness and outcome measurement are becoming embedded in the work that members choose to present, echoing the formalised research priorities of oncology nursing internationally (Rosenzweig et al., 2024).

Specialist competencies, advanced roles, and research infrastructure emerge as a shared axis of development. Topic 4 foregrounds professional roles, leadership, and organizational and ethical values, and Topic 3 foregrounds research networks and capacity. This aligns with the international expansion and diversification of advanced practice nursing roles in cancer care (Dowling et al., 2024), and it implies that role definition, leadership development, and research infrastructure are perceived as preconditions, not by-products, of clinical innovation.

Patient-reported outcomes and experiences, self-care, and therapeutic education are a second axis. Topic 2 (adherence, self-management, surveillance) and the measurement vocabulary of Topic 3 reflect growing attention to what patients report and to how nurses monitor it. This resonates with evidence that systematic, nurse-supported monitoring of patient-reported outcomes can improve outcomes in metastatic cancer (Basch et al., 2022), and it positions patient-reported outcome and experience measures (PROMs/PREMs), self-management support and therapeutic education as practical levers for continuity of care.

Palliative and supportive care and continuity (Topic 5) bridge clinical and organizational concerns, with case management and continuity recurring across topics. Nurse-led primary palliative care interventions have demonstrated value in advanced cancer (Schenker et al., 2021), and the corpus suggests that Italian oncology nurses are actively developing such models, including territorial and hospital–community integration.

Digital health and emerging technologies appear as an opportunity threaded through several topics rather than as a self-contained theme. The breadth of digital health and telehealth applications in cancer care is well documented (Shaffer et al., 2023), and telehealth-delivered models can match in-person care for some palliative endpoints (Greer et al., 2024). The relatively diffuse presence of digital language in the corpus may signal both genuine uptake and a risk of an organizational and competency divide if technological adoption outpaces role design, training, and infrastructure.

For AIIAO, these results offer an evidence-informed basis for agenda-setting. They suggest concrete directions: consolidating methodological and research-capacity training (continuing education aligned with measurement and study design); formalising advanced and specialist roles and the competencies that underpin safe care pathways; embedding PROMs/PREMs and self-management support into models of continuity; and developing a considered position on digital health that protects equity and competency. Because the latent topics cut across the editorial clusters, AIIAO might also reconsider how it structures abstract categories so that they better reflect the vocabulary members actually use.

Several limitations warrant discussion. The corpus reflects what was submitted to and included in a congress and is shaped by selection, language, contribution type, and participants' interests; the topics describe this corpus and do not establish the real-world prevalence of these phenomena in Italian oncology nursing. Topics are latent, overlapping dimensions, and dominant-topic assignment simplifies abstracts that often span several themes. Results from text-mining are sensitive to preprocessing and to the choice of k; we have therefore documented every decision and provided the full pipeline for audit. Therefore, the main limitations are the non-probabilistic nature of the corpus, heterogeneous abstract quality and length, residual linguistic noise from a community writing in both Italian and English, the sensitivity of results to

analytic choices, and the exploratory status of the topics, which should not be interpreted as a representative map of Italian oncology nursing.

Future work could extend this approach longitudinally, comparing successive AIIAO congresses to track how topics rise and fall, triangulating topic models with manual thematic coding, and linking abstract topics to subsequent full publications to assess which congress themes translate into sustained research programs.

Conclusions

A lexicometric and topic-modelling analysis of the AIIAO 2026 Congress abstracts identified five emerging topics centred on safe and standardized care pathways, self-management and adherence, research capacity and outcome measurement, professional roles and leadership, and palliative and supportive care. Interpreted prudently, these topics can help orient AIIAO's scientific agenda, continuing education, competency development, and policy positioning, while digital health is best treated as a cross-cutting opportunity to be managed equitably. The findings characterise the submitted corpus and are offered as a reproducible observatory of professional priorities rather than as a representative account of the field.

References

- Arun, R., Suresh, V., Veni Madhavan, C. E., & Narasimha Murty, M. (2010). On finding the natural number of topics with latent Dirichlet allocation: Some observations. In M. J. Zaki, J. X. Yu, B. Ravindran, & V. Pudi (Eds.), *Advances in knowledge discovery and data mining* (Lecture Notes in Computer Science, Vol. 6118, pp. 391–402). Springer. https://doi.org/10.1007/978-3-642-13657-3_43
- Basch, E., Schrag, D., Henson, S., et al. (2022). Effect of electronic symptom monitoring on patient-reported outcomes among patients with metastatic cancer: A randomized clinical trial. *JAMA*, 327(24), 2413–2422. https://doi.org/10.1001/jama.2022.9265
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*, 72(7–9), 1775–1781. https://doi.org/10.1016/j.neucom.2008.06.011
- Danler, M., Hackl, W. O., Neururer, S. B., Huber, L., & Pfeifer, B. (2024). Visualizing nursing narratives: An evaluation of latent Dirichlet allocation topic modeling for care reports. *Studies in Health Technology and Informatics*, 316, 1709–1713. https://doi.org/10.3233/SHTI240752
- Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document Numérique*, 17(1), 61–84. https://doi.org/10.3166/DN.17.1.61-84
- Dowling, M., Pape, E., Geese, F., et al. (2024). Advanced practice nursing titles and roles in cancer care: A scoping review. *Seminars in Oncology Nursing*, 40(3), Article 151627. https://doi.org/10.1016/j.soncn.2024.151627
- Greer, J. A., Temel, J. S., El-Jawahri, A., et al. (2024). Telehealth vs in-person early palliative care for patients with advanced lung cancer: A multisite randomized clinical trial. *JAMA*, 332(14), 1153–1164. https://doi.org/10.1001/jama.2024.13964
- Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl. 1), 5228–5235. https://doi.org/10.1073/pnas.0307752101
- Kocarnik, J. M., Compton, K., Dean, F. E., et al. (2022). Cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life years for 29 cancer groups from 2010 to 2019: A systematic analysis for the Global Burden of Disease Study 2019. *JAMA Oncology*, 8(3), 420–444. https://doi.org/10.1001/jamaoncol.2021.6987
- Oh, W. O., Lee, E., Heo, Y. J., Jung, M. J., & Han, J. (2024). Understanding global research trends in the control and prevention of infectious diseases for children: Insights from text mining and topic modeling. *Journal of Nursing Scholarship*, 56(4), 606–620. https://doi.org/10.1111/jnu.12963
- Rosenzweig, M., Belcher, S. M., Braithwaite, L. E., et al. (2024). Research priorities of the Oncology Nursing Society: 2024–2027. *Oncology Nursing Forum*, 51(6), 502–515. https://doi.org/10.1188/24.ONF.502-515
- Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In Proceedings of the 8th ACM International Conference on Web Search and Data Mining (WSDM '15) (pp. 399–408). *Association for Computing Machinery*. https://doi.org/10.1145/2684822.2685324
- Schenker, Y., Althouse, A. D., Rosenzweig, M., et al. (2021). Effect of an oncology nurse-led primary palliative care intervention on patients with advanced cancer: The CONNECT cluster randomized clinical trial. *JAMA Internal Medicine*, 181(11), 1451–1460. https://doi.org/10.1001/jamainternmed.2021.5185
- Shaffer, K. M., Turner, K. L., Siwik, C., et al. (2023). Digital health and telehealth in cancer care: A scoping review of reviews. *The Lancet Digital Health*, 5(5), e316–e327. https://doi.org/10.1016/S2589-7500(23)00049-3