CI PP IV CD CONFERÊNCIA INTERNACIONAL DE POLÍTICAS PÚBLICAS E CIÊNCIA DE DADOS
TERRITÓRIOS E SOCIEDADES EM REDE

Natural language processing and colorectal cancer: automated systematic review of trends in prediction algorithms and main risk factors

Geovanna Sobral Bermejo[1], Eduardo Palhares Júnior[2], James Moraes de Almeida[3], Natasha Fioretto Aguero[3], Adriano Honorato de Souza[2], Wenndisson da Silva Souza[2], Alexandre Lopes Martiniano[2]

[1] Universidade Anhembi Morumbi, geovanna.uam.sjc@gmail.com
[2] Instituto Federal de Educação, Ciência e Tecnologia do Amazonas, eduardo.palharesjr@ifam.edu.br, adriano.honorato@ifam.edu.br, wenndisson.souza@ifam.edu.br, alexandre.martiniano@ifam.edu.br
[3] Ilum Escola de Ciência, Centro Nacional de Pesquisa em Energia e Materiais, james.almeida@ilum.cnpem.br, natasha.aguero@cnpem.br

Abstract

The exponential growth of medical literature concerning prediction on Colorectal Cancer (CRC) hinders the selection of variables applicable to public healthcare systems. To investigate this gap, an analysis of technological evolution and clinical requirements of global AI research was proposed, based on the application of Natural Language Processing and Latent Dirichlet Allocation to a corpus of 2,383 articles published between 2015 and 2025. The results reveal a pronounced transition toward Deep Learning architectures and confirm that global models rely predominantly on structured variables, such as staging, demographics, and comorbidities. The conceptual mapping of these requirements against the Brazilian Ministry of Health's data dictionaries demonstrates that DataSUS possesses the necessary traceability and data maturity to support high-impact oncological predictors.

Keywords: *Colorectal Cancer; Natural Language Processing; Deep Learning; Topic Modeling; DataSUS.*

1. Introduction

Colorectal Cancer (CRC) remains one of the major challenges in contemporary oncology, ranking among the leading causes of morbidity and mortality globally (Siegel et al., 2020; Sung et al., 2021). Alongside its clinical severity, the scientific community confronts the phenomenon of Big Literature, an exponential growth in technical publications that renders the exhaustive synthesis of evidence for the development of new predictors virtually impossible (Obermeyer & Emanuel, 2016). Within this context, artificial intelligence, specifically through Natural Language Processing (NLP), emerges not merely as an extraction tool, but as a well-established methodological framework for mapping the knowledge architecture in oncology (Li et al., 2023; Topol, 2019; Yim et al., 2016).

Advances in computational methods have reconfigured predictive research in this field. Whereas classical Machine Learning algorithms dominated the literature over the last decade, there is now a transition toward Deep Learning architectures, driven by the need to process unstructured data and complex images (LeCun et al., 2015). Nevertheless, a critical gap persists between the development of these technologies in major research centers and their practical applicability in the healthcare systems of developing countries, such as Brazil.

The central problem addressed by this study is the difficulty in precisely identifying which clinical and sociodemographic variables are used in global predictive models, and whether these are represented in national databases. While cutting-edge global research frequently focuses on molecular biomarkers, the implementation of artificial intelligence in the Brazilian public healthcare system relies on the robustness of information systems such as DATASUS. Consequently, it is essential to validate whether the volume and quality of data generated by the national system meet the requirements to support these international models.

To this end, we conducted an automated systematic review using topic modeling based on LDA to analyze global prediction trends in CRC. In contrast to descriptive bibliometric reviews, this study focuses on mapping algorithmic requirements and analyzing variable usage patterns, culminating in a technical feasibility assessment of DataSUS. The ultimate objective is to provide an evidence-based roadmap for the development of predictive models that are simultaneously technologically robust and regionally applicable.

2. Methodology

The present study is based on an automated systematic review, structured using text mining and Natural Language Processing (NLP) techniques. The research pipeline was designed to extract and categorize predictive models and clinical variables directly from large volumes of scientific literature, using Python as its computational foundation.

2.1. Dataset Collection Strategy

To ensure the reproducibility and comprehensiveness of the search, data were extracted via the OpenAlex API, selected for its robustness as an open academic index (Priem et al., 2022). The standardized concept identifier for colorectal cancer (Concept ID: C526805850) was used, with a temporal window established from 2015 to the present to capture the period of greatest maturity in neural network architectures.

The extraction was refined using a mandatory Boolean filter ("artificial intelligence" OR "machine learning" OR "deep learning"), ensuring that the corpus was restricted to artificial

intelligence applications. The retrieved articles were mapped using a technical dictionary designed to classify the methods into five domains: (I) Deep Learning; (II) Ensemble/Boosting; (III) Classical Machine Learning; (IV) Natural Language Processing (NLP) Algorithms; and (V) Probabilistic Models. This process yielded a final dataset of 2,383 unique articles consistent with the scope of the study.

2.2. Natural Language Processing Pipeline

The textual analysis focused on publication abstracts. Preprocessing involved text normalization, tokenization, and lemmatization, using the spaCy library for its efficiency in high-performance natural language processing (Ines Montani et al., 2023). A customized list of medical stopwords was applied to eliminate generic oncology terms that could introduce statistical noise into the probabilistic models.

Knowledge structuring was carried out using the LDA, an unsupervised topic modeling technique that enables the identification of latent structures within document collections (Blei et al., 2003). The model was applied to the processed corpus to identify the five dominant themes in global research on CRC.

2.3. Mapping of Clinical Variables and Integration with DataSUS

In a second phase, a targeted LDA model for clinical variables was implemented. Instead of operating on the free text of the abstracts, this algorithm focused exclusively on a predefined dictionary of terms extracted from the corpus. This step aimed to isolate and cluster predictor variables — categorized into sociodemographic factors, diagnostic data, comorbidities, and healthcare infrastructure — alongside their respective oncological outcomes, such as mortality, costs, and length of hospital stay.

Finally, to assess the technical feasibility of enabling artificial intelligence models with Brazilian public data, the frequency and co-occurrence of these variables in the international literature were conceptually mapped against the information domains within DataSUS. Cross-referencing these textual data enabled the generation of visual maps that correlate the metric requirements of global models with the presence and structuring of these same data within the data infrastructure of the national healthcare system.

3. Results and Discussion

The results of this research are presented across three interconnected analytical axes: the temporal evolution of predictive technologies, the semantic mapping of the global literature, and the validation of their feasibility against data from the Brazilian healthcare system.

3.1. Technological and Thematic Evolution

Data extraction from the 2,383 articles reveals a profound methodological shift in predictive approaches in colorectal oncology. A chronological analysis of scientific output (**Figure 1**) indicates that, up to 2017, the literature relied almost exclusively on classical machine learning algorithms and ensemble models. From 2018 onward, a marked shift toward the widespread adoption of Deep Learning architectures is observed. This quantitative pattern in the dataset mirrors the theoretical consolidation of these technologies, as deep neural networks, propelled by the foundations established by LeCun et

al. (2015), rapidly became the dominant approach for processing complex oncological patterns (Pacal et al., 2020).

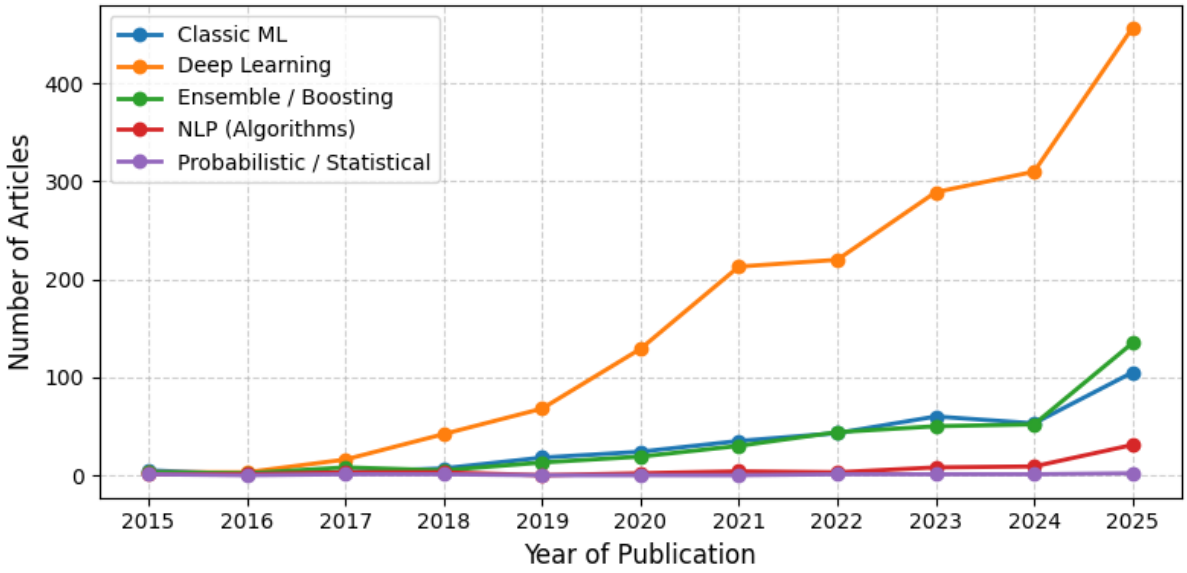

Figure 1: Temporal evolution of publication volume, highlighting the transition from Classical Machine Learning to Deep Learning architectures.

However, an analysis of algorithm families and their clinical application areas (**Figure 2**) reveals that this technological transition does not occur uniformly. The heatmap demonstrates that Deep Learning exhibits clear dominance in imaging and radiomics. Conversely, classical machine learning algorithms and ensemble methods, such as Random Forest and XGBoost, maintain a robust presence in the modeling of clinical data and genomics. As discussed by Cuocolo et al., (2020), this methodological persistence is explained by the structured and tabular nature of these data. In purely clinical predictive contexts, the explainability inherent to tree-based models often outweighs the predictive performance of deep learning architectures, which still face interpretability barriers.

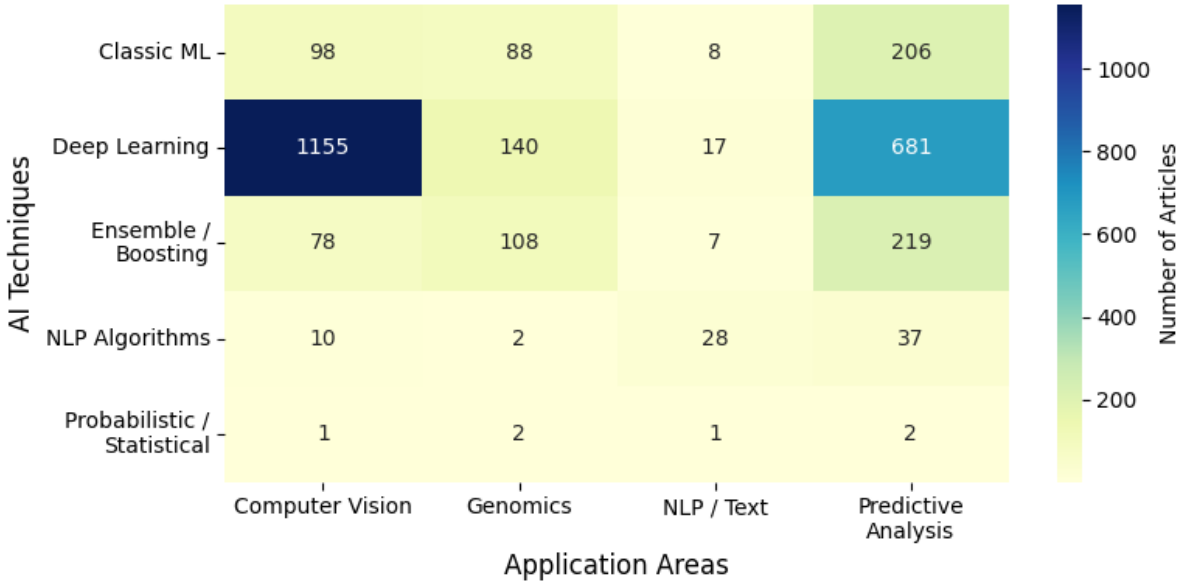

Figure 2: Heatmap comparing algorithm families with their primary clinical and molecular application areas.

To characterize the semantic structure underlying this massive volume of technological publications, the Latent Dirichlet Allocation model synthesized the textual corpus into five dominant research themes (**Figure 3**). The results indicate that global literature is strongly concentrated on early screening, particularly in screening and detection, followed by treatment response evaluation and precision medicine, particularly in genomics and biomarkers. This thematic configuration, as quantified by the model, is extensively corroborated by contemporary literature in colorectal oncology, which has prioritized the development of advanced diagnostic tools primarily focused on endoscopic imaging and computational pathology (Ahmad et al., 2021; Kudo et al., 2019; Skrede et al., 2020; Wang et al., 2019).

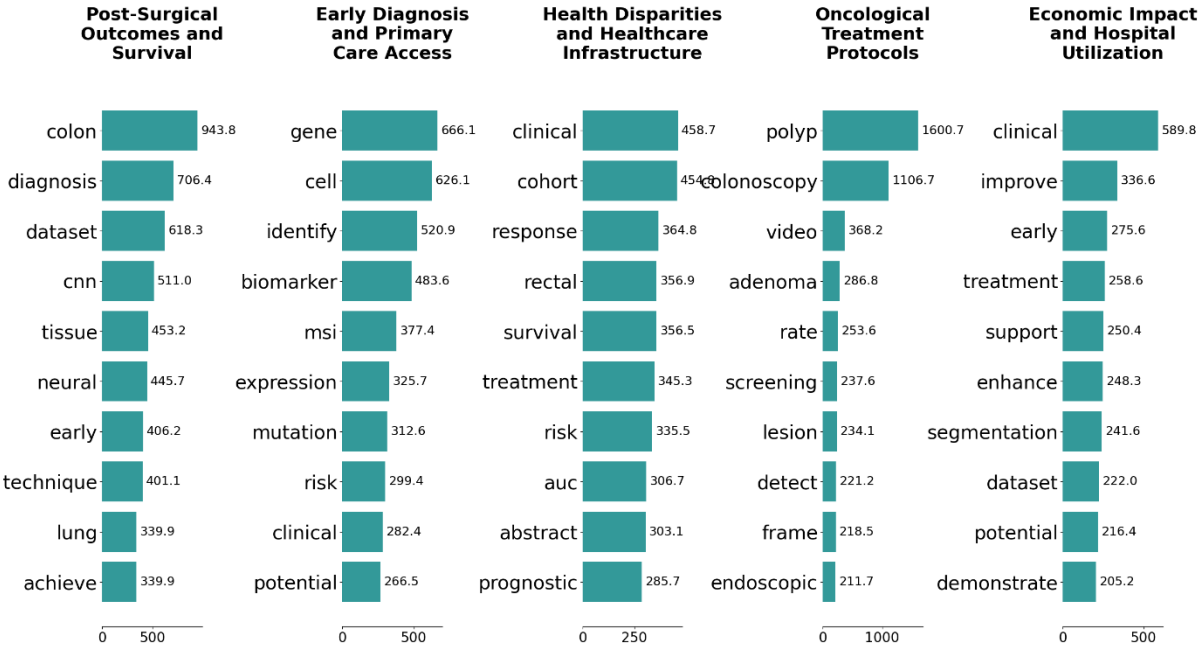

Figure 3: Distribution of the five primary thematic axes in the global literature on prediction in CRC, extracted via unsupervised topic modeling

This overview reinforces the notion that the oncological Big Literature phenomenon and the global state-of-the-art are strongly oriented toward the development of technologically complex diagnostic tools, such as radiomic and biomolecular features. This raises a critical question: does the public healthcare data infrastructure, particularly in developing countries, possess the required granularity to support and implement these global models? It is precisely this gap that the subsequent clinical modeling seeks to elucidate.

3.2. The Clinical Semantic Structure and Data Requirements

The identification of variables that underpin clinical prediction demands an analysis that transcends image-based approaches and focuses on tabular and structured data. To isolate the factors that underpin the models described in the literature, a clinical LDA model was applied once again, operating exclusively on the dictionary of predictive factors and outcomes. This process mapped the semantic relationships among variables (**Figure 4**), revealing that, when moving away from image-focused architectures, algorithms rely primarily on traditional epidemiological variables such as tumor staging (TNM), the presence

of comorbidities, and core demographic variables, including age, sex, and geographic location. This centrality of clinical records and diagnostic notes in the modeling of chronic diseases is corroborated by Sheikhalishahi et al. (2019), who highlight the potential of these data when processed using NLP techniques.

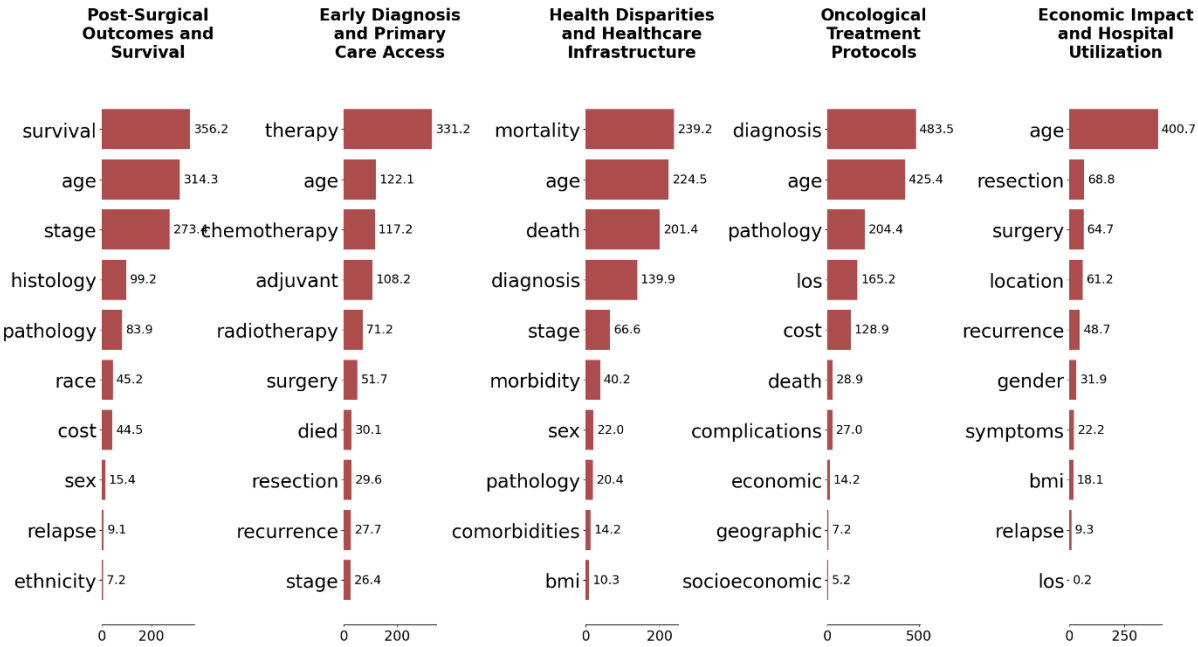

Figure 4: Latent topics extracted from the cross-referencing of predictors and outcomes in the global literature, evidencing the reliance on structured clinical variables.

The model's results demonstrate that the state-of-the-art in population-level prediction does not require the use of advanced molecular sequencing for each individual. Instead, the technical literature points to the need for a comprehensively populated sociodemographic and clinical data matrix that can be algorithmically integrated to predict critical outcomes, such as mortality, complication rates, and economic and institutional impact, measured in terms of costs and length of hospital stay. As discussed by Ahmad et al. (2021), although the future of pathology is promising in the biomolecular domain, the immediate impact of artificial intelligence in medical practice lies in its ability to transform routine clinical variables into high-performance decision-support tools.

3.3. Feasibility and Applicability in the Brazilian Context

The identification of key pillars of global predictive modeling establishes the methodological reference. The final stage of this study consisted of evaluating these international requirements against the Brazilian healthcare data infrastructure. The analysis of the temporal evolution and maturity of SUS databases (**Figure 5**) indicates continuous growth and stabilization of key variable collection over the past decade. This quantitative trend signals that the volume of records has already reached the critical mass required to meet the structural demands of machine learning algorithms.

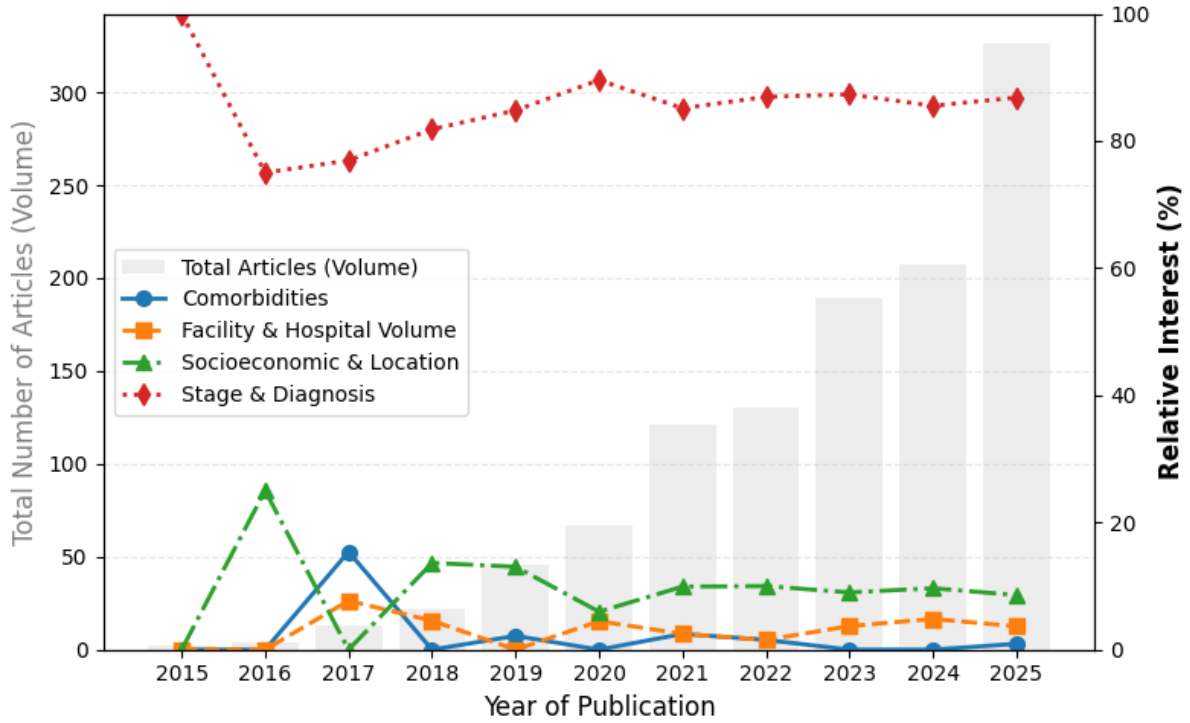

Figure 5: Temporal evolution of the volume of records for clinical and demographic variables in DataSUS.

To visualize the interaction of these variables throughout the patient journey, the flow diagram (**Figure 6**) illustrates the structural connectivity between the individual's baseline profile (demographics and initial diagnosis) and clinical outcomes. The mapping demonstrates a viable traceability structure that underpins the creation of consistent labels for training supervised models within the national context

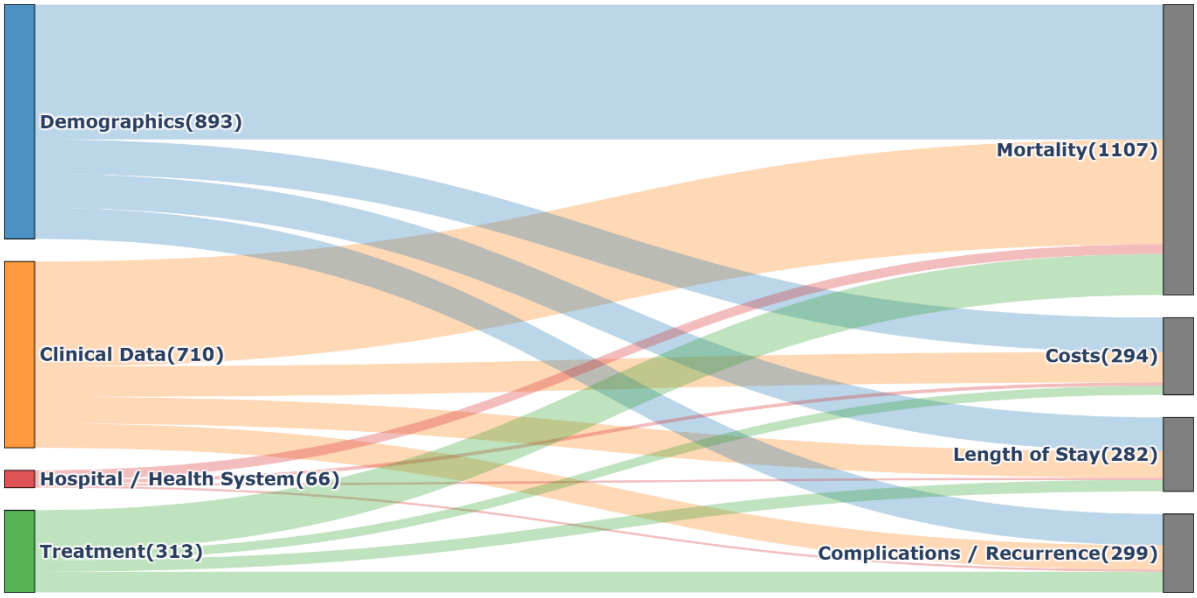

Figure 6: Sankey flow diagram illustrating the connectivity between baseline predictors and outcomes within the data ecosystem.

Technical feasibility is further supported by bivariate frequency analysis (**Figure 7**). This analysis directly correlates the academic relevance of each variable — reflected in its

prominence in the global literature — with its corresponding presence and structuring within the Ministry of Health's data dictionaries.

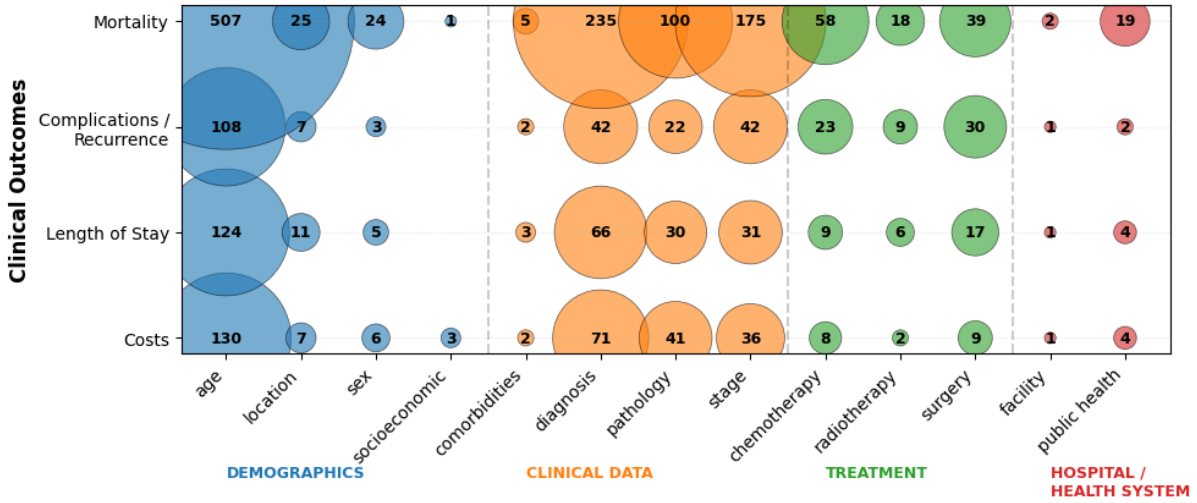

Figure 7: Scatter plot correlating the global literature's demand for predictive variables with their structural availability in Brazil

The factors with the greatest weight for risk prediction in CRC — namely age, staging, tumor location, adjuvant therapies, and comorbidities — lead the international literature and are well represented in Brazil. This alignment validates the central hypothesis of this study, as, despite historical fragmentation challenges, DataSUS constitutes a mature dataset capable of supporting advanced predictive technologies. Beyond mere infrastructural feasibility, the demographic richness of these databases addresses a pressing need in modern oncology. As argued by Srivastav et al. (2025), the mitigation of disparities and the effectiveness of artificial intelligence depend critically on the integration of social determinants of health. In this regard, the application of models to data from the Brazilian public system offers not only mathematical robustness but also a territorially responsive data framework capable of mitigating global technological bias.

4. Conclusion

The present study demonstrated the feasibility of Natural Language Processing and topic modeling for the systematic analysis of large-scale oncological literature. The analysis of 2,383 articles enabled not only the identification of the technological transition from classical machine learning to Deep Learning architectures, but also the characterization of the semantic structure that drives the global development of predictive algorithms for colorectal cancer (CRC).

The primary contribution of this research lies in the bridge established between the state-of-the-art in artificial intelligence and the realities of healthcare infrastructure in developing countries. The automated extraction revealed that, alongside advancements in molecular biomarkers, the international literature is heavily grounded in structured clinical predictors—such as staging, demographics, therapies, and comorbidities.

The conceptual mapping of these variables against the Ministry of Health's data dictionaries demonstrated that DataSUS possesses the density, traceability, and data completeness required to serve as a primary data source for training high-impact predictive models.

Regarding methodological limitations, the textual analysis was restricted to publication abstracts. Although this approach optimizes the computational processing of large data volumes and minimizes textual noise, it may omit secondary variables reported exclusively in full texts. Additionally, the reliance on English-language search terms may underrepresent epidemiological characteristics from region-specific publications.

As a natural progression, future work should advance from the data validation stage to practical implementation. Using the optimized set of variables identified in this study—such as age, tumor location, and history of comorbidities—subsequent research should focus on the training and calibration of a custom predictive algorithm. The ultimate objective is to develop a risk and survival prediction tool that, aligned with global technological standards, is fully compatible with and applicable to the operational reality of the Brazilian Unified Health System (SUS).

References

Ahmad, Z., Rahim, S., Zubair, M., & Abdul-Ghafar, J. (2021). Artificial intelligence (AI) in medicine, current applications and future role with special emphasis on its potential and promise in pathology: Present and future impact, obstacles including costs and acceptance among pathologists, practical and philosophical considerations. A comprehensive review. *Diagnostic Pathology*, *16*(1), 24. https://doi.org/10.1186/s13000-021-01085-4

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Cuocolo, R., Caruso, M., Perillo, T., Ugga, L., & Petretta, M. (2020). Machine Learning in oncology: A clinical appraisal. *Cancer Letters*, *481*, 55–62. https://doi.org/10.1016/j.canlet.2020.03.032

Ines Montani, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, & Henning Peters. (2023). *explosion/spaCy: V3.7.2: Fixes for APIs and requirements* (Version v3.7.2) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.1212303

Kudo, S., Mori, Y., Misawa, M., Takeda, K., Kudo, T., Itoh, H., Oda, M., & Mori, K. (2019). Artificial intelligence and colonoscopy: Current status and future perspectives. *Digestive Endoscopy*, *31*(4), 363–371. https://doi.org/10.1111/den.13340

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Li, C., Zhang, Y., Weng, Y., Wang, B., & Li, Z. (2023). Natural Language Processing Applications for Computer-Aided Diagnosis in Oncology. *Diagnostics*, *13*(2), 286. https://doi.org/10.3390/diagnostics13020286

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the Future—Big Data, Machine Learning, and Clinical Medicine. *New England Journal of Medicine*, *375*(13), 1216–1219. https://doi.org/10.1056/NEJMp1606181

Pacal, I., Karaboga, D., Basturk, A., Akay, B., & Nalbantoglu, U. (2020). A comprehensive review of deep learning in colon cancer. *Computers in Biology and Medicine*, *126*, 104003. https://doi.org/10.1016/j.compbiomed.2020.104003

Priem, J., Piwowar, H., & Orr, R. (2022). *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2205.01833

Sheikhalishahi, S., Miotto, R., Dudley, J. T., Lavelli, A., Rinaldi, F., & Osmani, V. (2019). Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. *JMIR Medical Informatics*, *7*(2), e12239. https://doi.org/10.2196/12239

Siegel, R. L., Miller, K. D., Goding Sauer, A., Fedewa, S. A., Butterly, L. F., Anderson, J. C., Cercek, A., Smith, R. A., & Jemal, A. (2020). Colorectal cancer statistics, 2020. *CA: A Cancer Journal for Clinicians*, *70*(3), 145–164. https://doi.org/10.3322/caac.21601

Skrede, O.-J., De Raedt, S., Kleppe, A., Hveem, T. S., Liestøl, K., Maddison, J., Askautrud, H. A., Pradhan, M., Nesheim, J. A., Albregtsen, F., Farstad, I. N., Domingo, E., Church, D. N., Nesbakken, A., Shepherd, N. A., Tomlinson, I., Kerr, R., Novelli, M., Kerr, D. J., & Danielsen, H. E. (2020). Deep learning for prediction of colorectal cancer outcome: A discovery and validation study. *The Lancet*, *395*(10221), 350–360. https://doi.org/10.1016/S0140-6736(19)32998-8

Srivastav, A. K., Singh, A., Singh, S., Rivers, B., Lillard, J. W., & Singh, R. (2025). Revolutionizing Oncology Through AI: Addressing Cancer Disparities by Improving Screening, Treatment, and Survival Outcomes via Integration of Social Determinants of Health. *Cancers*, *17*(17), 2866. https://doi.org/10.3390/cancers17172866

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, *71*(3), 209–249. https://doi.org/10.3322/caac.21660

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, *25*(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

Wang, P., Berzin, T. M., Glissen Brown, J. R., Bharadwaj, S., Becq, A., Xiao, X., Liu, P., Li, L., Song, Y., Zhang, D., Li, Y., Xu, G., Tu, M., & Liu, X. (2019). Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: A prospective randomised controlled study. *Gut*, *68*(10), 1813–1819. https://doi.org/10.1136/gutjnl-2018-317500

Yim, W., Yetisgen, M., Harris, W. P., & Kwan, S. W. (2016). Natural Language Processing in Oncology: A Review. *JAMA Oncology*, *2*(6), 797. https://doi.org/10.1001/jamaoncol.2016.0213