CIPP
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, Eduardo Palhares Júnior, James Moraes de Almeida, Natasha Fioretto Aguero, Adriano Honorato de Souza, Wenndisson da Silva Souza, Alexandre Lopes Martiniano

ipb INSTITUTO POLITÉCNICO DE BRAGANÇA Escola Superior de Comunicação, Administração e Turismo
CITeD CENTRO INVESTIGAÇÃO TRANSDISCIPLINAR EM eDUCAÇÃO E DESENVOLVIMENTO
UNIAG Applied Management Research Unit
RP3CD
fct Fundação para a Ciência e a Tecnologia

Colorectal Cancer: A Global Public Health Challenge

- 3rd most common cancer worldwide and 2nd leading cause of cancer-related death
- Over 1.9 million new cases and 900,000 deaths estimated globally in 2020
- Rising incidence in younger populations and in low- and middle-income countries
- Early-stage detection dramatically improves survival rates, yet diagnosis remains late in most public health systems

The Challenge of Big Literature in Oncology

- The exponential growth of medical publications makes exhaustive manual review virtually impossible
- In the CRC + AI intersection alone: thousands of articles published between 2015 and 2025
- Traditional systematic reviews are slow, expensive, and inevitably incomplete
- A new methodological challenge: how to synthesize this volume without losing scientific rigor?

The Central Research Question

Global AI research on CRC is advancing rapidly, but which clinical variables do these models actually rely on?

Are those variables available in public health databases of developing countries, specifically, in Brazil's DataSUS?

Can NLP itself be the tool to answer these questions at scale?

Why OpenAlex?

- Fully open platform — no licensing, no access barriers, replicable by any public institution
- Covers 250M+ scholarly works with structured metadata including abstracts, citations and hierarchical concept tags
- Unlike PubMed (10k record API limit) or Scopus (proprietary), OpenAlex enables unrestricted large-scale extraction
- Methodological coherence: a study concluding that public data infrastructure is viable must itself rely on open infrastructure

Concept-Based Corpus Identification

What is the official CRC concept ID today?

- Search by name
- API returns valid ID **`C526805850`**
- 666,872 articles

How to reliably intersect CRC with AI?

- `.search("machine learning")`
- 6,859 results
- captures indirect mentions

Does this ID work as a filter?

- Apply ID + year filter
- 119,247 articles (2023–2024 only)
- Concept is active and at scale

- `.filter(title_and_abstract)`
- 2,841 results
- restricted to title and abstract

Two Extraction Strategies, One Corpus

Application Area

- Computer Vision
- Predictive Analysis
- NLP Text Mining
- Genomics

59,959 unique articles

Captures articles where AI appears as context

Algorithm Family

- Classic ML
- Deep learning
- Ensemble / Boosting
- NLP Algorithms
- Probabilistic / Statistical

2,383 unique articles

Captures articles where AI is the method

Text Preprocessing: Why a Medical NLP Model?

spaCy NLP Framework

The spaCy is a industrial-grade Natural Language Processing architecture applied for automated normalization, tokenization, and linguistic analysis of large-scale scientific texts.

`en_core_sci_lg` Model

- Biomedical Specialization: Pre-trained specifically on a massive scientific corpus, unlike generic NLP models.
- Semantic Accuracy: Ensures correct identification and extraction of complex clinical entities that standard models misclassify.
- Noise Reduction: Applies context-aware stopword filtering to eliminate non-informative medical jargon, directly optimizing the accuracy of the downstream LDA topic modeling.

LDA 1: Global Thematic Analysis

Strategic Objective: Unsupervised Topic Modeling to map the architecture of knowledge in the global Colorectal Cancer and AI research landscape.

Pipeline

- Standardization and Tokenization
- Lemmatization
- Stopwords Removal
- Vectorization
- Topic Modeling

Key Insight: 5 latent topics extracted from the corpus detecting technological shifts and tematic concentration

LDA 2: Clinical Requirements & DataSUS Compatibility

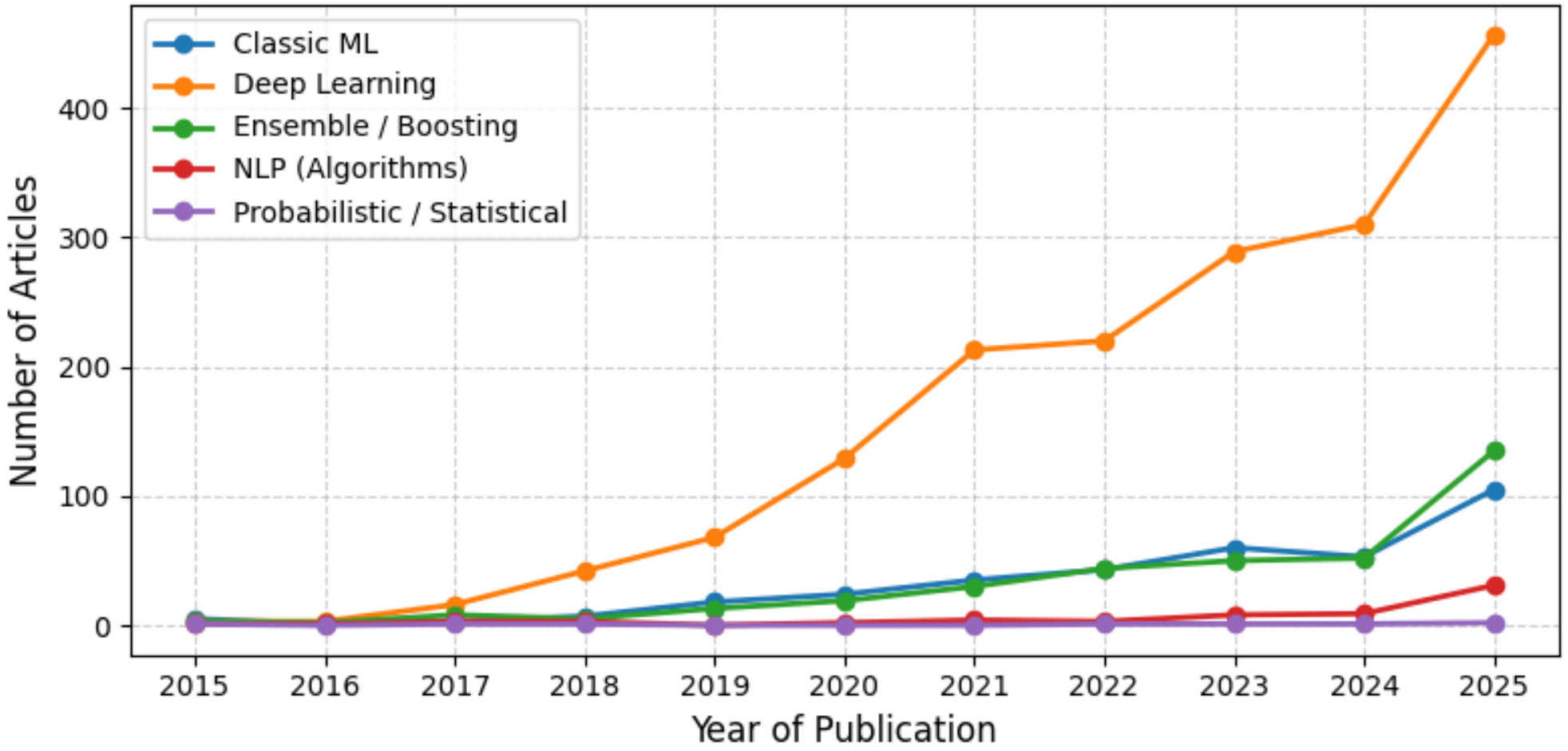
Strategic Objective: Identify the specific clinical and demographic variables required by global AI models to validate their availability and maturity within the Brazilian Public Health System (DataSUS).

A restricted vocabulary was used to map clinical variables (predictors and outcomes) aligned to the DataSUS categories.

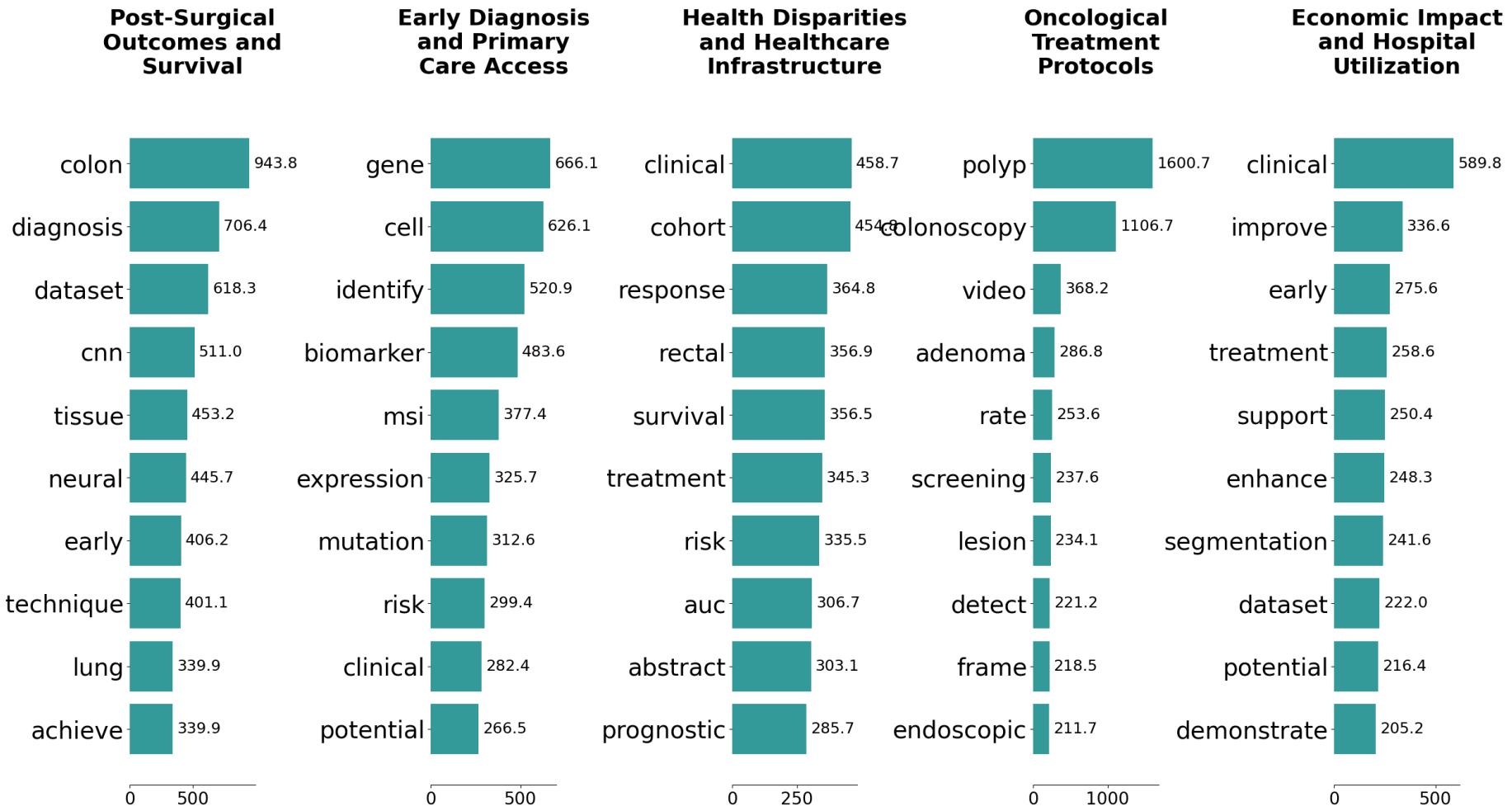
Pipeline

- Controlled Dictionary
- Targeted Extraction
- Binary Vectorization
- Clustering

From Classical ML to Deep Learning: A Decade of Transition



What Is the Global CRC AI Literature Actually Discussing?

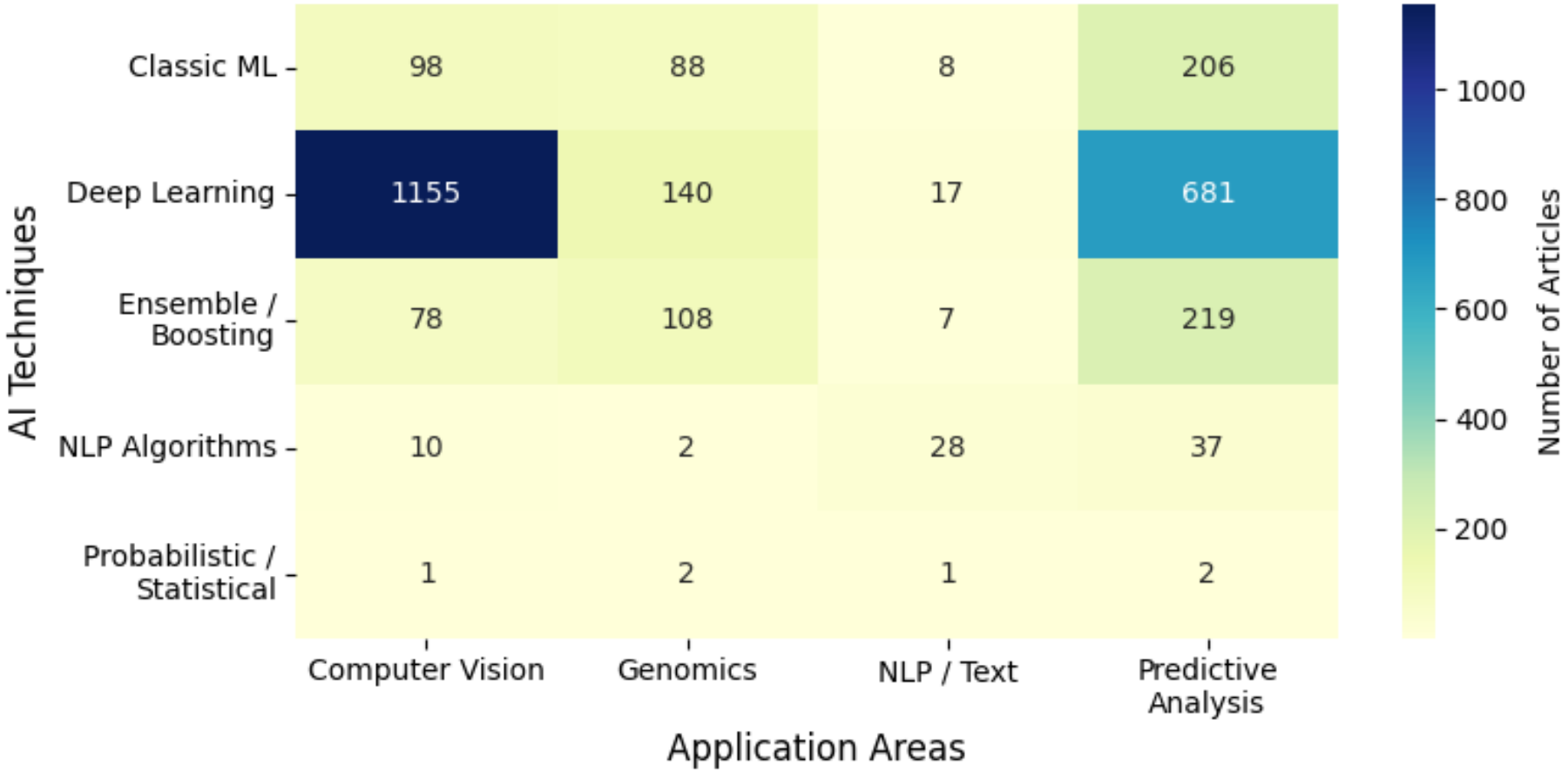


Five Dominant Themes — And What They Reveal

- **Public Health, Screening and Prevention** — early detection dominates the global agenda
- **Digital Pathology, Slide Analysis and Histology** — computational pathology is a major research frontier
- **Clinical Management, Therapies and Treatment Response** — AI applied to treatment decision support
- **Prognosis, Survival and Risk Factors** — predicting outcomes from clinical variables
- **Classification and Deep Learning Architectures** — the methodological engine behind all other themes

Global research is heavily concentrated on technologically complex diagnostic tools, which raises the central question: does public health data infrastructure in developing countries have the granularity to support these models?

AI Techniques vs. Clinical Application Areas

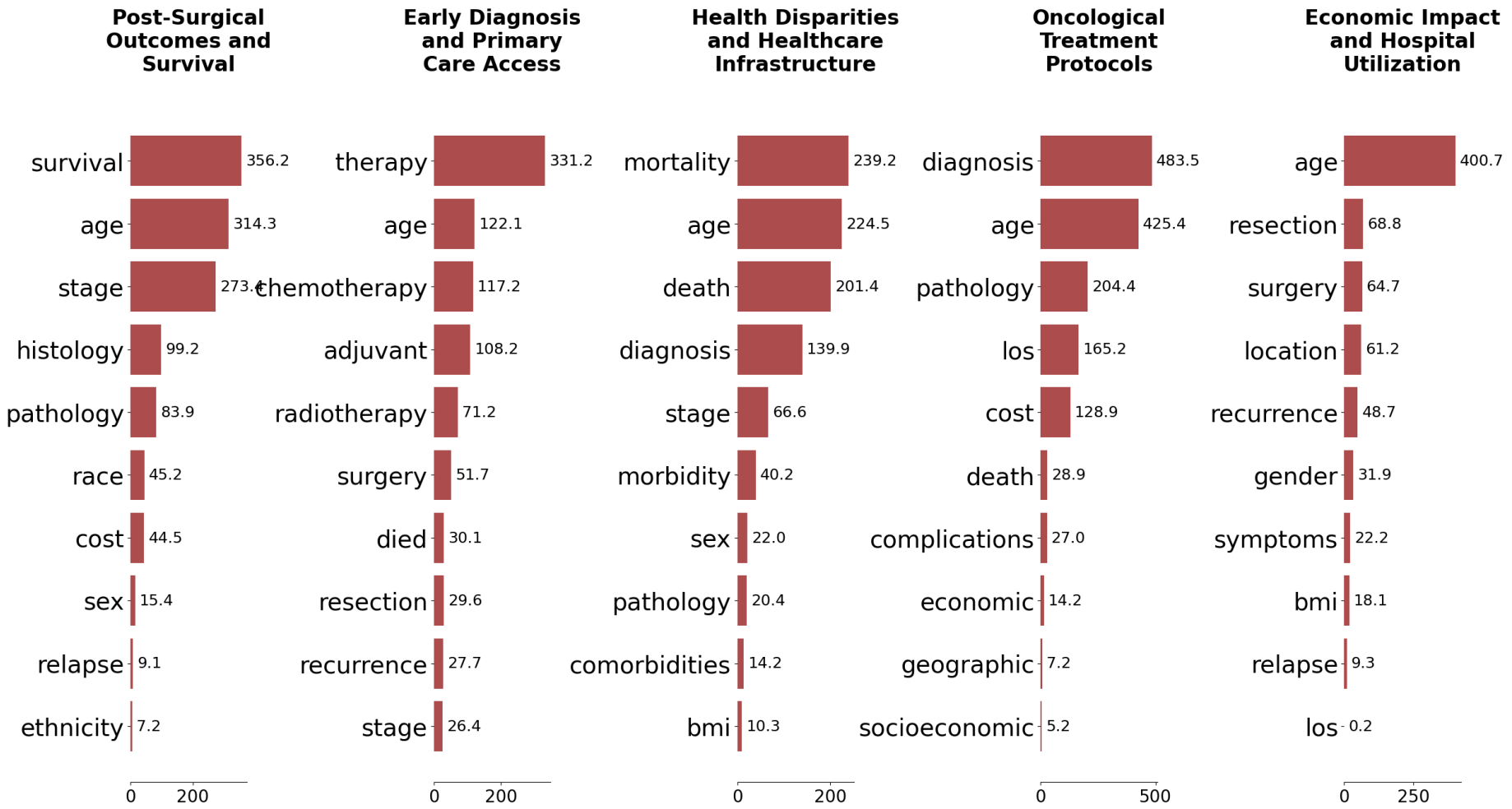


The Nuance Behind the Transition

- Deep Learning dominates **Computer Vision** — imaging, endoscopy, histopathology
- Classical ML and Ensemble methods (Random Forest, XGBoost) maintain strong presence in **clinical and genomic data**
- In purely clinical predictive contexts, **explainability outweighs raw performance** — tree-based models remain preferred
- The technological transition is real, but it is **not uniform across application areas**

The most policy-relevant finding: advanced imaging infrastructure is not a prerequisite to deploy AI in oncology — structured clinical data is sufficient to build high-impact predictive model

What Do Global Predictive Models Actually Require?



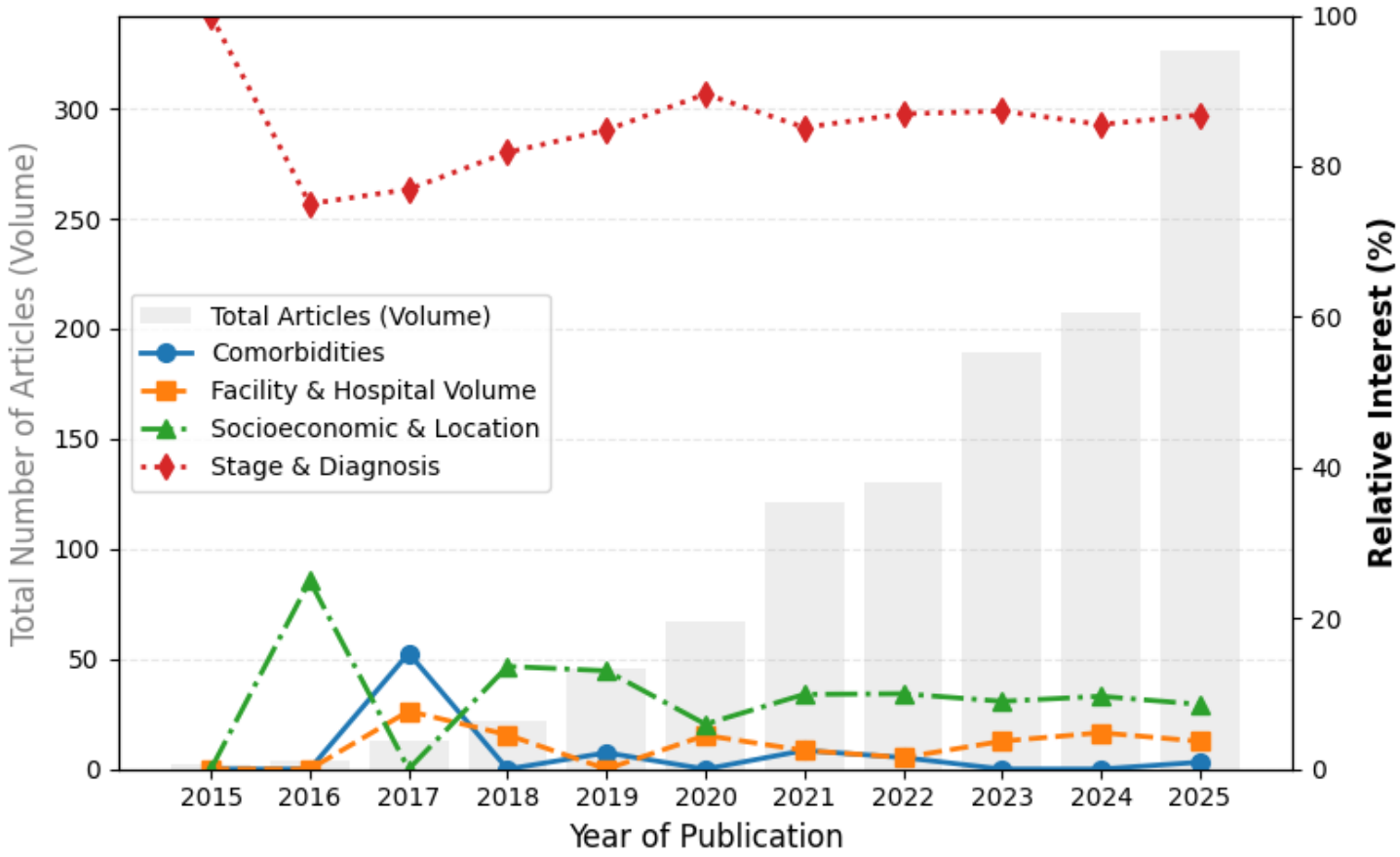
Structured Clinical Variables Drive Global Predictive Models

- **Post-Surgical Outcomes and Survival** — age, staging, histology and pathology dominate survival prediction
- **Early Diagnosis and Primary Care Access** — therapy type, adjuvant treatment and recurrence as key predictors
- **Health Disparities and Healthcare Infrastructure** — mortality, morbidity and comorbidities as population-level indicators
- **Oncological Treatment Protocols** — diagnosis, pathology and complications driving treatment modeling
- **Economic Impact and Hospital Utilization** — age, resection, surgery and location predicting institutional costs

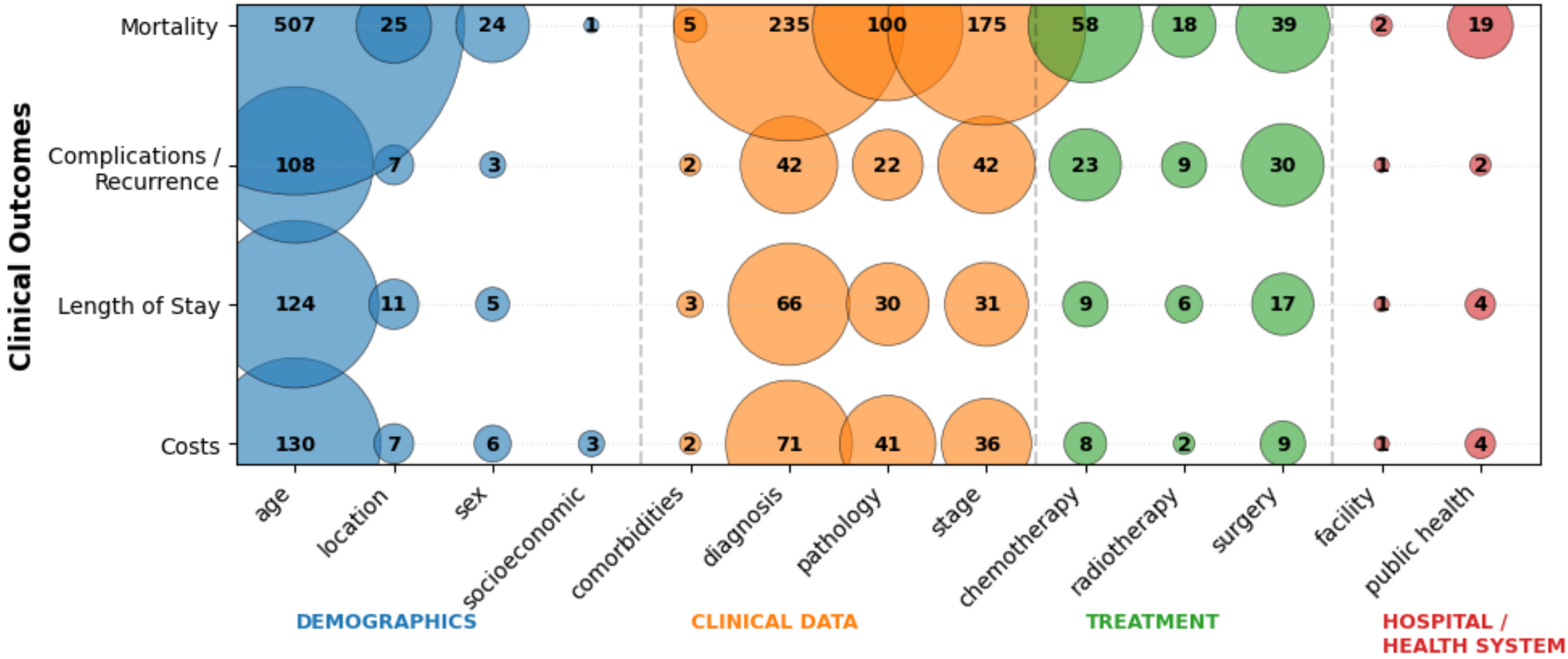
State-of-the-art population-level prediction does not require molecular sequencing — it requires a well-populated sociodemographic and clinical data matrix

DataSUS Variable Collection: A Decade of Growth and Stabilization

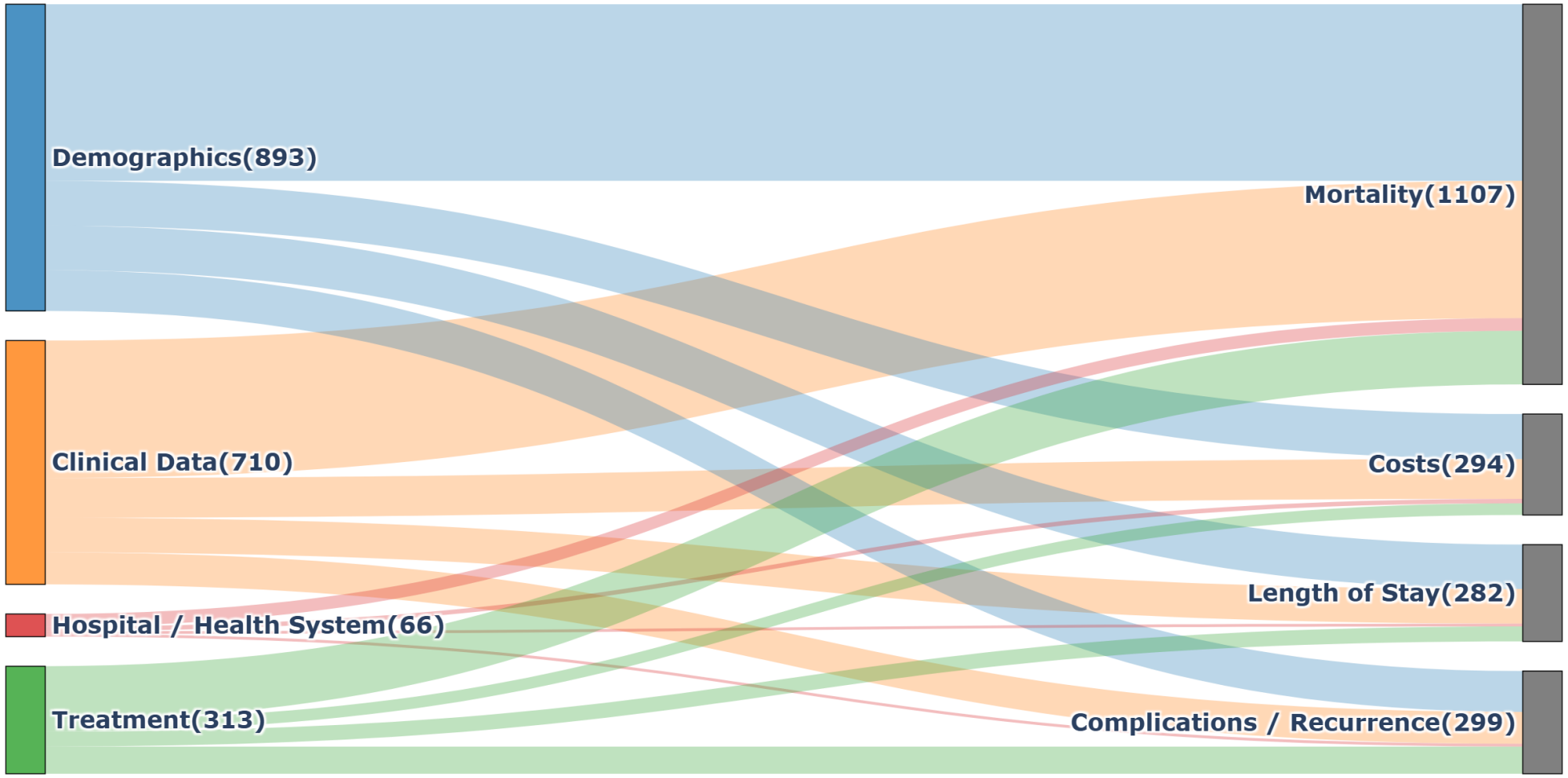
- The number of articles increases every year, but relative interest shows consolidation
- Stage & Diagnosis is the subject of greatest research interest
- Volume of records has reached the critical mass required by machine learning algorithms
- Temporal stability and data maturity are verified



Global Demand vs. Brazilian Data Availability



The Patient Journey: From Baseline Predictors to Clinical Outcomes

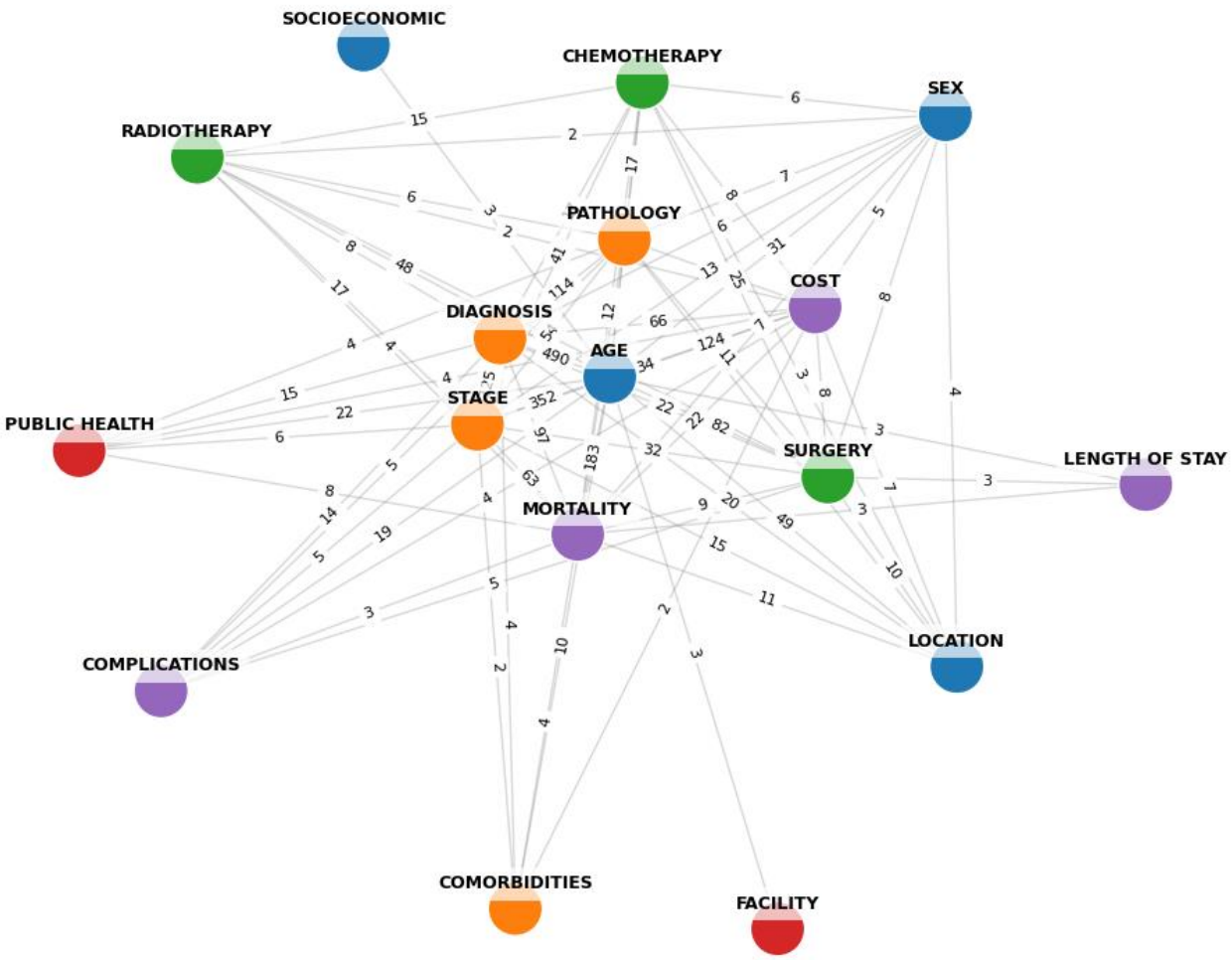


Integrated Interpretation: Bubble Plot and Sankey Diagram

- Bubble Plot identifies the most frequent predictor groups and clinical outcomes in the global literature.
- Sankey Diagram shows how these predictor groups connect to each outcome across the predictive pipeline.
- Demographics and Clinical Data are the main input categories, and Mortality is the most frequently predicted outcome.
- Together, these figures demonstrate that structured clinical and sociodemographic variables form the core of population-level predictive models.
- This evidence confirms that DataSUS contains the essential variables required to develop AI models compatible with international practice.

Variable Co-occurrence Network in CRC AI Research

- Global AI research is centered on a stable clinical core, indicating that these variables are consistently modeled together.
- Less connected variables, such as socioeconomic and health system factors, remain underexplored.
- These peripheral variables represent promising opportunities for future predictive models using DataSUS.



From Evidence to Action: A Roadmap for AI in Brazilian Oncology

- Natural Language Processing and Topic Modeling enabled a scalable and reproducible systematic review of global AI research in Colorectal Cancer.
- Structured clinical and sociodemographic variables remain the foundation of population-level predictive modeling.
- DataSUS contains the essential variables required to develop predictive models aligned with international standards.
- Brazil already possesses the data infrastructure needed to implement AI-driven oncology within the Sistema Único de Saúde.

Future Research Directions

- Develop and validate predictive models for mortality and survival using DataSUS.
- Compare machine learning and deep learning approaches on Brazilian real-world data.
- Incorporate underexplored variables, including socioeconomic and health system factors.
- Design an interpretable clinical decision-support tool for oncology in the Sistema Único de Saúde.

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

Thanks for Watching

INSTITUTO POLITÉCNICO DE BRAGANÇA ipb Escola Superior de Comunicação, Administração e Turismo

CITeD CENTRO INVESTIGAÇÃO TRANSDISCIPLINAR EM eDUCAÇÃO E DESENVOLVIMENTO

UNIAG Applied Management Research Unit

RP3CD

fct Fundação para a Ciência e a Tecnologia