Preliminary report

What Primary Care Records Reveal About Cancer Signs and Symptoms

Exploratory Analysis for Early Detection conducted with health data from the City of Recife (2016–2024)

Authors: Arthur Lorenzi, Daniela Krausz, Luciana Vasconcelos Sardinha, and Pedro de Paula, from Vital Strategies, in collaboration with Prof. Dr. Tiago Torrent from the FrameNet Brasil Laboratory at the Federal University of Juiz de Fora (UFJF), and Ronaldo Corrêa Ferreira da Silva, an oncologist at the National Cancer Institute (INCA) who contributed as an independent consultant to this study, and with Bruna dos Passos Gimenes, Aline Leal Gonçalves Creder Lopes and Amanda Maria Chaves from the Secretariat of Specialized Care at the Ministry of Health.

Executive Summary

In Brazil, approximately 60% of cancer cases are diagnosed at an advanced stage. Early identification is one of the most effective strategies to reverse this trend. Detecting signs and symptoms before formal diagnosis allows treatment to begin in a timely manner, enabling less invasive interventions and higher therapeutic success rates. Primary Health Care (PHC) medical records within the Unified Health System (SUS) contain rich clinical information about these signs, recorded at each consultation over the years. Despite this, this source remains largely underused for that purpose.

This exploratory analysis, conducted in Recife with data from 2016 to 2024, investigated four priority cancer types: breast, cervical, prostate, and colorectal. The results indicate that signs related to these cancers are already recorded by health professionals in PHC records months or years before formal diagnosis. These signs represent a relevant source of clinical information that conventional ICD-10 coding cannot fully capture. One of the central challenges is making them visible and transforming them into qualified, reliable, and actionable information. To this end, Vital Strategies Brasil developed a semantic analysis methodology that processes free-text fields in medical records and systematically identifies these signs.

Key Findings

- 425 distinct clinical expressions were identified in medical records prior to diagnosis, demonstrating the richness of free text as a clinical information source.
- For cervical cancer, semantic analysis identified 21.8 times more people with early signs than conventional ICD code search.
- For colorectal cancer, zero signs were identified via ICD codes, but around 33% of patients had signs recorded in free text.
- The first sign appears, at the median, 9 months before a breast cancer diagnosis and 19 months before a cervical cancer diagnosis.
- The analyzable universe is 3.4 times larger than what was explored in this analysis: as new data sources are integrated, the potential grows substantially.

Background

Cancer is one of the leading causes of death in Brazil and, in some municipalities, is already the leading cause of death from disease. To reduce mortality, two strategies are fundamental: reducing incidence through risk factor control, and identifying cases early. Although reducing risk factors has greater potential impact, its implementation is more complex and results occur over the long term. Early detection aims to identify cancer at the first signs and symptoms, when treatment can be most effective and least aggressive, and includes not only early diagnosis but also prompt initiation of treatment, whose delay can undermine all the benefits achieved.

In Brazil, approximately 60% of cancer cases are diagnosed at an advanced stage, which helps explain the persistence of high mortality rates. Countries that achieved significant reductions have invested in PHC-based early diagnosis programmes, identifying warning signs and referring suspected cases swiftly. In Brazil, clinical data recorded in PHC remains largely underused as a source of information for this purpose.

What We Did

This exploratory analysis was conducted in Recife, Pernambuco, Brazil, based on health data accessed by Vital Strategies Brasil under a Technical Cooperation Agreement with the municipality, covering the period from 2016 to 2024. The focus is on four priority cancer types: breast, cervical, prostate, and colorectal, selected for their incidence, mortality, and the expected presence of recordable signs in PHC before formal diagnosis.

How Cases Were Identified

To identify patients with a cancer diagnosis and trace their history in primary health care, four complementary data sources were used:

- PEC e-SUS PHC: longitudinal clinical history with free-text medical records from 2016 to 2024
- RCBP (Population-Based Cancer Registry): the only source with a formal diagnosis date per patient, serving as the central anchor of the analysis. Data covering the period 2022 to 2024
- SIM (Mortality Information System): cases recorded at the time of death between 2016 and 2024
- SIH (Hospital Information System): cases recorded at hospital admissions from 2016 to 2022 (cases between 2022–2024 are included in the RCBP)

The sources were cross-referenced through data linkage using a deterministic algorithm, reconstructing each patient's trajectory in primary health care prior to diagnosis. The Recife RCBP available for this analysis covers only 2022 to 2024 and is in the process of consolidation, which restricts the universe of patients with a validated diagnosis date and characterises the results as preliminary.

The analysis starts from 2,865 patients with a confirmed diagnosis from the RCBP. Of these, 1,229 (42.9%) had attended PHC before diagnosis, generating 13,573 pre-diagnosis consultations, the central basis of this analysis. This corresponds to an average of 11.0 consultations per patient over the pre-diagnosis period, with a median of 7.0.

When hospital admission records (SIH) for the period 2016 to 2022 and mortality records (SIM) are integrated, the universe of identified patients rises to 9,829 people. The 5,589 people identified exclusively through SIM do not have a validated diagnosis date, which prevents temporal analysis. As new sources are integrated, the volume of analysable cases will grow significantly.

Coverage varies by cancer type. Cervical (69.1%) and breast (39.6%) cancers show greater PHC presence, reflecting the role of primary health care in identifying suspicious signs and symptoms and referring cases for diagnostic investigation. In addition, these cancers have well-established clinical guidelines and recommendations for periodic screening, which strengthens their detection within PHC. For prostate and colorectal cancers, the lower coverage may reflect both the age profile of the affected population, which tends to use PHC less frequently, and the absence of national guidelines directing the screening of these cancers in primary health care.

How Signs Were Identified

The identification of early signs in medical records combines two approaches:

ICD-10 code search: checks whether a consultation contains coded diagnoses associated with each cancer type. Fast and structured, but limited to what the health professional chose to code.

Semantic analysis of free text: uses an artificial intelligence model developed by Vital Strategies Brasil in partnership with FrameNet Brasil (UFJF), trained on health language data in Portuguese. The model processes the open fields of medical records, what the patient reports and what the professional records, and identifies clinical expressions associated with early signs of cancer, regardless of vocabulary variations, regionalisms, or abbreviations.

This combination makes it possible to capture signs that exist in the data but that conventional coding does not record.

What We Found

1. What Conventional Coding Misses and Semantic Analysis Reveals

The most striking result of this analysis is the difference between what ICD code searches capture and what semantic analysis makes visible:

Cancer type	Via ICD (people)	Via semantics (people)	Multiplier
Prostate (N=50)	28 (56.0%)	29 (58.0%)	1.0x
Colorectal (N=55)	0 (0.0%)	18 (32.7%)	N/A
Breast (N=484)	63 (13.0%)	209 (43.2%)	3.3x
Cervical (N=700)	13 (1.9%)	284 (40.6%)	218x

The table shows that, for most cancers analysed, conventional ICD-10 coding captures only a fraction of the signs present in medical records. For cervical cancer, ICD-10 identified 13 people with recorded signs; the semantic analysis found 284. For colorectal cancer, ICD-10 identified no one; the semantic analysis found 18. For breast cancer, ICD-10 captured fewer than one third of the people identified by semantic analysis. Only for prostate cancer did both methods yield similar results. These signs were recorded in the medical records; semantic analysis is what makes them visible. It is not a substitute for ICD-10: it is complementary, enabling greater use of information already present in the record, making visible what conventional coding does not capture.

2. Signs Identified With Significant Lead Time

For breast and cervical cancers, the cancer types with sufficient volume for temporal analysis, the distance between the first identified sign and formal diagnosis was calculated. Prostate (N=5) and colorectal (N=18) were omitted due to insufficient volume for robust conclusions.

Cancer type	Median (days)	Median (approx. months)	N
Cervical	581 days	~19 months	284
Breast	271 days	~9 months	209

For cervical cancer, 44.4% of people had their first sign more than 2 years before diagnosis. For breast cancer, 36.5% had their first sign within the window of less than 6 months before diagnosis. These figures represent lower bounds: future iterations with an expanded semantic vocabulary will tend to identify more people and with greater lead time.

Next Steps

The results of this analysis point to concrete pathways for subsequent steps, both in refining the identified signs and in expanding the analyzable dataset.

Refining the Identified Signs

The current analysis maps the universe of signs without differentiating by degree of specificity. The next step is to differentiate signs by degree of known clinical association with each cancer type, analysing first those with the strongest association, then those with medium and low association. It is also important to calculate the positive predictive value per sign and time window, to assess the extent to which the lead time of a sign is associated with its specificity, and to select the signs with the greatest potential for early identification in PHC.

Expanding the Analyzable Dataset

The Recife Population-Based Cancer Registry available for this analysis covers only 2022 to 2024, restricting the universe of patients with a validated diagnosis date. Expansion to a complete RCBP base in Recife and at national scale will substantially increase the analytical potential. In addition, other sources that provide diagnosis dates, such as SIH, SIA, Siscan, and RHC, may be considered to expand case coverage beyond the RCBP period.

Expanding and Refining the Semantic Vocabulary

The current semantic vocabulary for early signs, though comprehensive, can still be expanded. Future iterations will seek to increase the set of identified clinical expressions and semantic patterns, including composite patterns such as signs associated with specific anatomical locations, with the expectation of increasing both the volume of detected signs and the lead time of detection.

As these efforts advance, the potential for early cancer detection from primary care data will be assessed with greater precision and robustness.

The ultimate goal is to build a replicable method, scalable to other cancer types and contexts, capable of supporting health professionals and managers in identifying signs and symptoms suggestive of cancer, contributing to the timely identification and referral of suspected cases.

Technical Annex

Databases and Record Linkage

The analysis used four complementary health data sources from Recife, accessed through a Technical Cooperation Agreement:

- PEC e-SUS PHC: primary health care consultations with free-text medical records (SOAP fields)
- Population-Based Cancer Registry: Population-Based Cancer Registry, collected by INCA, covering 2022–2024. The only source with a formal diagnosis date per patient
- SIM: Mortality Information System
- SIH: Hospital Information System

Since the Population-Based Cancer Registry includes pre-consolidation records, deterministic record linkage was performed for deduplication and cross-source integration. All numbers reported in this document are per patient-cancer: a patient with two distinct cancer types is counted twice. When a patient had multiple records in the Population-Based Cancer Registry, the most detailed or most recent ICD code was retained.

Examples of ICD-10 Codes Used

Cancer type	No. of codes	Example
Prostate	15	R31 — Unspecified hematuria
Colorectal	14	D50.0 — Iron deficiency anemia secondary to blood loss
Breast	12	N63 — Unspecified breast lump
Cervical	15	N95.0 — Postmenopausal bleeding
Total	56	—

Examples of Semantic Terms Searched

Cancer type	No. of terms	Examples
Prostate	23	dysuria, nocturia, urinary retention, bladder fullness
Colorectal	25	constipation, rectal bleeding, rectocolitis
Breast	19	dysplasia, mastodynia, mastalgia
Cervical	50	dysmenorrhea, dyspareunia, menorrhagia
Total	117	—

Examples of Semantic Patterns

#	Description	Cancer(s)	Examples
1	Symptom/condition in the breast	Breast	breast cyst, nipple lesion
2	Symptom/condition in the pelvic/ uterine region	Cervical	endometrial inflammation, vaginal hemorrhage

#	Description	Cancer(s)	Examples
3	Symptom/condition in the lower limbs	Cervical	leg swelling
4	Symptom/condition in the intestinal/rectal region	Colorectal	intestinal ulcer, rectal polyp
5	Symptom/condition in the abdominal region	Colorectal	abdominal distension, lower abdominal pain
6	Symptom/condition in the lumbar region	Prostate / Cervical	lower back pain
7	Symptom/condition in the perineal region	Prostate / Cervical	perineal pain
8	Bleeding or pain when urinating	Prostate	pain when urinating
9	Bleeding or pain when defecating	Colorectal	intestinal hemorrhage, pain when defecating
10	Pain during sexual intercourse	Cervical	pain during intercourse

Glossary

PHC: Primary Health Care. The first level of care in the Brazilian Unified Health System (SUS), serving as the entry point to the health system.

Semantic analysis: a natural language processing technique that identifies the meaning of words and expressions in context, not just their presence.

Term bank: a set of clinical expressions associated with early signs of each cancer type, generated during the semantic structure modeling phase.

ICD-10: International Classification of Diseases, 10th edition. Standardized diagnostic coding system used in medical records.

Clinical expressions: forms of recording symptoms or health conditions identified in free-text fields of medical records.

Longitudinal clinical history: the set of visits and records for the same patient over time.

Linkage / Record linkage: the process of identifying and integrating records belonging to the same patient across different databases.

Population-Based Cancer Registry: Population-Based Cancer Registry. A database that consolidates information on cancer cases diagnosed in a population, collected for INCA.

Sign / Symptom / Health condition: clinical manifestations that may precede a formal diagnosis. In this document, the term "sign" is used broadly to include symptoms reported by the patient and conditions recorded by the professional.

Positive Predictive Value (PPV): the proportion of people with a given sign who actually developed the disease. A measure of diagnostic specificity.

Published in May 2026