Relatório preliminar

O que os prontuários da atenção primária revelam sobre sinais e sintomas de câncer

Análise exploratória para identificação precoce, conduzida com dados de saúde do Município de Recife (2016-2024)

Autores: Arthur Lorenzi, Daniela Krausz, Luciana Vasconcelos Sardinha, Pedro de Paula, da Vital Strategies, com colaboração do Prof. Dr. Tiago Torrent, do Laboratório FrameNet Brasil da Universidade Federal de Juiz de Fora (UFJF), do Ronaldo Corrêa Ferreira da Silva, Médico Oncologista que atua no Instituto Nacional do Câncer e contribuiu como consultor independente no âmbito deste estudo, da Bruna dos Passos Gimenes, Aline Leal Gonçalves Creder Lopes e Amanda Maria Chaves da Secretaria de Atenção Especializada no Ministério da Saúde.

Sumário Executivo

No Brasil, cerca de 60% dos casos de câncer são diagnosticados em estágio avançado. A identificação precoce é uma das estratégias mais eficazes para reverter esse quadro. Detectar sinais e sintomas antes do diagnóstico formal permite que o tratamento seja iniciado em tempo oportuno,possibilitando intervenções menos invasivas e maiores taxas de sucesso terapêutico. Os prontuários da Atenção Primária (APS) no Sistema Único de Saúde (SUS) concentram informações clínicas ricas sobre esses sinais, registradas a cada atendimento ao longo dos anos. Apesar disso, essa fonte permanece amplamente subutilizada para essa finalidade.

Esta análise exploratória, conduzida em Recife com dados de 2016 a 2024, investigou quatro tipos de câncer prioritários: mama, colo do útero, próstata e colorretal. Os resultados indicam que sinais relacionados a esses cânceres já estão registrados por profissionais de saúde nos prontuários da APS, com meses ou anos de antecedência em relação ao diagnóstico formal. Esses sinais representam uma fonte relevante de informação clínica que a codificação convencional por CID não consegue capturar integralmente. Um dos desafios centrais é torná-los visíveis e transformá-los em informações qualificadas, confiáveis e acionáveis. Para isso, a Vital Strategies Brasil desenvolveu uma metodologia de análise semântica que processa os campos de texto livre dos prontuários e identifica esses sinais de forma sistemática.

Principais achados

- 425 expressões clínicas distintas foram identificadas nos prontuários antes dos diagnósticos, evidência da riqueza do texto livre como fonte de informação clínica.
- Para o câncer de colo do útero, a análise semântica identificou 21,8 vezes mais pessoas com sinais precoces do que a busca convencional por código CID.
- Para o câncer colorretal, zero sinais foram identificados via CID, mas cerca de 33% das pessoas tinham sinais registrados em texto livre.
- O primeiro sinal aparece, em mediana, 9 meses antes do diagnóstico de mama e 19 meses antes do diagnóstico de colo do útero.
- O universo analisável é 3,4 vezes maior do que o explorado nesta análise: à medida que novas fontes forem integradas, o potencial cresce de forma expressiva.

Contexto

O câncer é uma das principais causas de morte no Brasil e, em alguns municípios, já é a maior causa de morte por doença. Para reduzir a mortalidade, duas estratégias são fundamentais: reduzir a incidência, por meio do controle dos fatores de risco, e identificar os casos precocemente. Embora a redução dos fatores de risco tenha maior potencial de impacto, sua implementação é mais complexa e os resultados ocorrem a longo prazo. A detecção precoce tem como objetivo identificar o câncer nos primeiros sinais e sintomas, quando o tratamento pode ser mais eficaz e menos agressivo, e inclui não apenas o diagnóstico precoce, mas o início rápido do tratamento, cujo atraso pode comprometer todo o benefício obtido.

No Brasil, cerca de 60% dos casos de câncer são diagnosticados em estágio avançado, o que ajuda a explicar a persistência de altas taxas de mortalidade. Países que obtiveram reduções expressivas investiram em programas de diagnóstico precoce baseados na atenção primária, identificando sinais de alerta e encaminhando casos suspeitos rapidamente. No Brasil, os dados clínicos registrados na APS permanecem amplamente subutilizados como fonte de informação para essa finalidade.

O que fizemos

Esta análise exploratória foi conduzida em Recife, Pernambuco, Brasil, com base nos dados de saúde acessados pela Vital Strategies Brasil via Acordo de Cooperação Técnica com o município, cobrindo o período de 2016 a 2024. O foco está em quatro tipos de câncer prioritários: mama, colo do útero, próstata e colorretal, selecionados por sua incidência, mortalidade e pela presença esperada de sinais registráveis na APS antes do diagnóstico formal.

Como os casos foram identificados

Para identificar pacientes com diagnóstico de câncer e rastrear seu histórico na atenção primária, foram utilizadas quatro fontes de dados complementares:

- PEC e-SUS APS: histórico clínico longitudinal com prontuários em texto livre de 2016 a 2024
- RCBP (Registro de Câncer de Base Populacional): única fonte com data formal de diagnóstico por paciente, sendo a âncora central da análise. Dados do período de 2022 a 2024
- SIM (Sistema de Informação sobre Mortalidade): casos registrados no momento do óbito entre 2016 e 2024
- SIH (Sistema de Informações Hospitalares): casos registrados em internações no período de 2016 a 2022 (Os casos entre 2022-2024 estão inclusos no RCBP)

As fontes foram cruzadas por meio de pareamento de dados (linkage) com algoritmo determinístico, reconstruindo a trajetória de cada paciente na atenção primária antes do diagnóstico. O RCBP de Recife disponível para esta análise cobre apenas 2022 a 2024 e está em fase de consolidação, o que restringe o universo de pacientes com data de diagnóstico validada e caracteriza os resultados como preliminares.

A análise parte de 2.865 pacientes com diagnóstico confirmado pelo RCBP. Desses, 1.229 (42,9%) passaram pela APS antes do diagnóstico, gerando 13.573 atendimentos pré-diagnóstico, a base central desta análise. Isso corresponde a uma média de 11,0 atendimentos por paciente ao longo do período pré-diagnóstico, com mediana de 7,0.

Quando se integram registros de internação (SIH) no período de 2016 a 2022 e óbito (SIM), o universo de pacientes identificados sobe para 9.829 pessoas. As 5.589 pessoas identificadas exclusivamente por SIM não possuem data de diagnóstico validada, o que impede a análise temporal. À medida que novas fontes forem integradas, o volume de casos analisáveis crescerá de forma expressiva.

A cobertura varia por tipo de câncer. Colo do útero (69,1%) e mama (39,6%) apresentam maior presença na APS, reflexo do papel da atenção primária na identificação de sinais e sintomas suspeitos e no encaminhamento para investigação diagnóstica. Além disso, esses cânceres possuem diretrizes clínicas consolidadas e recomendação de exames periódicos para rastreamento, o que fortalece sua detecção no âmbito da APS. Para próstata e colorretal, a cobertura menor pode refletir tanto o perfil etário da população acometida, que tende a frequentar menos a APS, quanto a ausência de diretrizes nacionais que orientem o rastreamento desses cânceres na atenção primária.

Como os sinais foram identificados

A identificação de sinais precoces nos prontuários combina duas abordagens:

Busca por código CID-10: verifica se o atendimento contém diagnósticos codificados associados a cada tipo de câncer. Rápida e estruturada, mas limitada ao que o profissional de saúde escolheu codificar.

Análise semântica do texto livre: utiliza um modelo de inteligência artificial desenvolvido pela Vital Strategies Brasil em parceria com a FrameNet Brasil (UFJF), treinado em dados linguísticos de saúde em português. O modelo processa os campos abertos dos prontuários, o que o paciente relata e o que o profissional registra, e identifica expressões clínicas associadas a sinais precoces de câncer, independentemente de variações de vocabulário, regionalismos ou abreviaturas.

Essa combinação permite capturar sinais que existem nos dados mas que a codificação convencional não registra.

O que encontramos

1. O que a codificação convencional não captura e a análise semântica revela

O resultado mais expressivo desta análise é a diferença entre o que a busca por código CID captura e o que a análise semântica torna visível:

Tipo de câncer	Via CID (pessoas)	Via semântica (pessoas)	Multiplicador
Próstata (N=50)	28 (56,0%)	29 (58,0%)	1,0x
Colorretal (N=55)	0 (0,0%)	18 (32,7%)	s/d
Mama (N=484)	63 (13,0%)	209 (43,2%)	3,3x
Colo do Útero (N=700)	13 (1,9%)	284 (40,6%)	21,8x

A tabela evidencia que, para a maioria dos cânceres analisados, a codificação convencional por CID captura apenas uma parcela dos sinais presentes nos prontuários. Para o câncer de colo do útero, o CID identificou 13 pessoas com sinais registrados; a análise semântica encontrou 284. Para o colorretal, o CID não identificou nenhuma pessoa; a análise semântica encontrou 18. Para mama, o CID capturou menos de um terço das pessoas identificadas pela análise semântica. Apenas para próstata os dois métodos chegaram a resultados parecidos.

Esses sinais estavam registrados nos prontuários; a análise semântica é o que os torna visíveis. Ela não é substituta do CID: é complementar, permitindo fazer uso de mais informação já presente no prontuário, tornando visível o que a codificação convencional não captura.

2. Sinais com antecedência significativa

Para mama e colo do útero, tipos de câncer com volume suficiente para análise temporal, foi calculada a distância entre o primeiro sinal identificado e o diagnóstico formal. Próstata (N=5) e colorretal (N=18) foram omitidos por não possuírem volume suficiente para conclusões robustas.

Tipo de câncer	Mediana (dias)	Mediana (meses aprox.)	N
Colo do Útero	581 dias	~19 meses	284
Mama	271 dias	~9 meses	209

Para o colo do útero, 44,4% das pessoas têm o primeiro sinal com mais de 2 anos de antecedência. Para mama, 36,5% estão na janela de menos de 6 meses antes do diagnóstico. Esses números representam limites inferiores: futuras iterações com vocabulário semântico expandido tenderão a identificar mais pessoas e com maior antecedência

Próximos passos

Os resultados desta análise apontam caminhos concretos para as etapas seguintes, tanto na qualificação dos sinais identificados quanto na expansão da base analisável.

Qualificação dos sinais identificados

A análise atual mapeia o universo de sinais sem diferenciar por grau de especificidade. A próxima etapa é diferenciar os sinais por grau de associação clínica conhecida com cada tipo de câncer, analisando primeiro os de maior associação, depois os de média e baixa. Também é importante calcular o valor preditivo positivo por sinal e janela temporal, avaliar em que medida a antecedência do sinal está associada à sua especificidade, e selecionar os sinais com maior potencial de identificação precoce na APS.

Expansão da base analisável

O RCBP de Recife disponível para esta análise cobre apenas 2022 a 2024, restringindo o universo de pacientes com data de diagnóstico validada. A expansão para uma base completa do RCBP em Recife e em escala nacional ampliará substancialmente o potencial analítico. Além disso, outras bases que fornecem data de diagnóstico, como SIH, SIA, Siscan e RHC, podem ser consideradas para ampliar a cobertura de casos fora do período do RCBP.

Expansão e refinamento do vocabulário semântico

O vocabulário semântico atual referentes a sinais precoces, apesar de abrangente, ainda pode ser expandido. As próximas iterações buscarão aumentar o conjunto de expressões clínicas e padrões semânticos identificados, incluindo padrões compostos como sinais associados a localizações anatômicas específicas, com expectativa de aumento no volume de sinais detectados e na antecedência de detecção.

À medida que essas frentes avançarem, o potencial de identificação precoce de câncer a partir dos dados da APS poderá ser dimensionado com maior precisão e robustez.

O objetivo final é construir um método replicável, passível de expansão para outros tipos de câncer e contextos, capaz de apoiar profissionais e gestores de saúde na identificação de sinais e sintomas sugestivos de câncer, contribuindo para a identificação e encaminhamento oportuno dos casos suspeitos.

Anexo Técnico

Bancos de dados e linkage

A análise utilizou quatro fontes complementares de dados de saúde de Recife, acessadas via Acordo de Cooperação Técnica:

- PEC e-SUS APS:atendimentos da atenção primária com prontuários em texto livre (campos SOAP)
- RCBP: Registro de Câncer de Base Populacional, coletado pelo INCA, cobrindo 2022–2024. Única fonte com data formal de diagnóstico por paciente
- SIM: Sistema de Informação sobre Mortalidade
- SIH: Sistema de Informações Hospitalares

Como o RCBP inclui registros pré-consolidação, foi realizado pareamento de dados com algoritmo determinístico para deduplicação e integração entre fontes. Todos os números reportados neste documento são por paciente-câncer: um paciente com dois tipos de câncer distintos é contado duas vezes. Quando um paciente possuía múltiplos registros no RCBP, foi mantido o código CID mais detalhado ou o mais recente.

Exemplos de códigos CID-10 utilizados

Tipo de câncer	Nº códigos	Exemplo
Próstata	15	R31 — Hematúria não especificada
Colorretal	14	D50.0 — Anemia por deficiência de ferro secundária à perda de sangue
Mama	12	N63 — Nódulo mamário não especificado
Colo do Útero	15	N95.0 — Sangramento pós-menopausa
Total	56	—

Exemplos de termos semânticos buscados

Tipo de câncer	Nº termos	Exemplos
Próstata	23	disúria, nictúria, retenção urinária, bexigoma
Colorretal	25	obstipação, enterorragia, retocolite
Mama	19	displasia, mastodinia, mastalgia
Colo do Útero	50	dismenorreia, dispareunia, menorragia
Total	117	—

Exemplos de padrões semânticos

#	Descrição	Câncer(es)	Exemplos
1	Sintoma/condição na mama	Mama	cisto no seio, ferida no mamilo
2	Sintoma/condição na região pélvica/uterina	Colo do útero	inflamação do endométrio, hemorragia vaginal
3	Sintoma/condição nos membros inferiores	Colo do útero	inchaço das pernas

#	Descrição	Câncer(es)	Exemplos
4	Sintoma/condição na região intestinal/reto	Colorretal	úlcera intestinal, pólipo no reto
5	Sintoma/condição na região abdominal	Colorretal	distensão abdominal, dor no baixo ventre
6	Sintoma/condição na região lombar	Próstata / Colo do útero	dor na lombar
7	Sintoma/condição na região perineal	Próstata / Colo do útero	algia na região perineal
8	Sangramento ou dor ao urinar	Próstata	dor ao urinar
9	Sangramento ou dor ao evacuar	Colorretal	hemorragia intestinal, dor ao evacuar
10	Dor na relação sexual	Colo do útero	dor ao se relacionar

Glossário

APS: Atenção Primária à Saúde. Primeiro nível de atenção do SUS, porta de entrada do sistema de saúde.

Análise semântica: técnica de processamento de linguagem natural que identifica o significado de palavras e expressões em contexto, não apenas sua presença.

Banco de termos: conjunto de expressões clínicas associadas a sinais precoces de cada tipo de câncer, gerado durante a fase de modelagem das estruturas semânticas.

CID-10: Classificação Internacional de Doenças, 10ª edição. Sistema padronizado de codificação diagnóstica utilizado nos prontuários.

Expressões clínicas: formas de registro de sintomas ou condições de saúde identificadas nos campos de texto livre dos prontuários.

Histórico clínico longitudinal: conjunto de atendimentos e registros de um mesmo paciente ao longo do tempo.

Linkage / Pareamento de dados: processo de identificação e integração de registros pertencentes ao mesmo paciente em diferentes bases de dados.

RCBP: Registro de Câncer de Base Populacional. Base de dados que consolida informações sobre casos de câncer diagnosticados em uma população, coletada para o INCA.

Sinal / Sintoma / Condição de saúde: manifestações clínicas que podem preceder um diagnóstico formal. Neste documento, o termo "sinal" é usado de forma ampla para incluir sintomas relatados pelo paciente e condições registradas pelo profissional.

Valor Preditivo Positivo (VPP): proporção de pessoas com um determinado sinal que efetivamente desenvolveram a doença. Medida de especificidade diagnóstica.

Publicado em maio de 2026