Preparação e Integração de Dados da Atenção Primária e Vigilância da Tuberculose para Suporte ao Treinamento de Modelos de Linguagem

Samantha Mendonça[1,2], **Sandro Moreira**[3], **Daniel Sacramento**[1,3], **Daniel Castro**[1,2,4], **Vanderson Sampaio**[1,2,5]

[1]Escola Superior de Ciências da Saúde – Universidade do Estado do Amazonas (UEA)
Manaus – AM – Brasil

[2]Fundação de Medicina Tropical Dr. Heitor Vieira Dourado (FMT-HVD)
Manaus – AM – Brasil

[3]Fundação Oswaldo Cruz – Instituto Leônidas e Maria Deane (ILMD)
Manaus – AM – Brasil

[4]Instituto Nacional de Pesquisas da Amazônia (INPA) – Manaus – AM – Brasil

[5]Instituto Todos pela Saúde (ITpS)
São Paulo – SP – Brasil

{samanthaalecrim.biotec, sandrofmstm, dsacramento.am, danielbarrosbio, vandersons}@gmail.com

Abstract. *The scarcity of integrated clinical data limits the development of Large Language Models (LLMs) to support decision-making in public health. This work presents a data engineering and probabilistic matching pipeline to integrate records from the Primary Care (PEC) and the Tuberculosis Surveillance System (SINAN-TB) in Manaus. From a dataset of 7.8 million raw records, the process resulted in 1.5 million unique individuals and 8,596 matched records with high confidence (score $\geq$ 0.90). This unified database serves as the foundation for training models capable of identifying suspected cases in clinical narratives, strengthening surveillance and diagnostic support for tropical diseases.*

Resumo. *A escassez de dados clínicos integrados limita o desenvolvimento de Modelos de Linguagem de Grande Escala (LLMs) para suporte à decisão na saúde pública. Este trabalho apresenta um pipeline de engenharia de dados e pareamento probabilístico para integrar registros do Prontuário Eletrônico do Cidadão (PEC) e do SINAN-TB em Manaus. A partir de 7,8 milhões de registros brutos, o processo resultou em 1,5 milhão de indivíduos únicos e 8.596 registros pareados com alta confiança ($\geq$ 0,90). Esta base unificada é o alicerce para o treinamento de modelos capazes de identificar casos suspeitos em narrativas clínicas, fortalecendo a vigilância e o suporte ao diagnóstico de doenças tropicais.*

1. Introdução

A detecção precoce e a notificação oportuna de doenças transmissíveis são componentes centrais da vigilância epidemiológica, pois condicionam a interrupção de cadeias de transmissão e a efetividade das respostas em saúde pública. Esses processos, entretanto, dependem do reconhecimento inicial de casos suspeitos no ponto de cuidado. Em contextos marcados por elevada demanda assistencial e fluxos de atenção complexos, esse reconhecimento pode ser tardio ou não ocorrer, comprometendo a oportunidade da intervenção e a qualidade da informação disponível para a vigilância [Meckawy et al. 2022].

Nesse contexto, narrativas clínicas da Atenção Primária à Saúde (APS) tornam-se fontes cruciais, capturando sinais de agravos antes de sua incorporação aos sistemas de notificação oficiais [Seinen et al. 2024]. A busca por termos sentinelas em prontuários eletrônicos tem demonstrado eficácia na detecção precoce de surtos, superando atrasos inerentes ao fluxo de digitação e disponibilidade de dados [Sampaio et al. 2023]. Embora o avanço dos prontuários eletrônicos e das técnicas de processamento de linguagem natural (PLN) amplie interesse por esse tipo de aplicação, a literatura aponta limitações por problemas recorrentes, como escassez de dados adequados, validação insuficiente, baixa padronização entre contextos institucionais e fraca integração com os fluxos dos sistemas de saúde [Hossain et al. 2023].

Esse ponto é particularmente crítico em agravos como a tuberculose, nos quais a informação se encontra fragmentada entre diferentes fontes clínicas, laboratoriais e de vigilância. Além disso, o cuidado do paciente, geralmente, é caracterizado por um longo período para o diagnóstico, seguimento prolongado, e múltiplos pontos de cuidado, exigindo forte articulação entre assistência e vigilância. Revisão de escopo recente mostrou que as soluções digitais aplicadas à tuberculose na APS abrangem desde apoio ao diagnóstico e notificação até monitoramento da adesão e gestão programática, evidenciando tanto o potencial quanto os desafios de integração dessas abordagens aos serviços de saúde [Sacramento et al. 2026]. Por outro lado, estudos com bases relacionadas à doença no Brasil descreveram fragilidades na identificação de casos, incompletude em variáveis de seguimento e ganhos analíticos decorrentes do relacionamento entre sistemas [Moreira et al. 2025][Lima et al. 2020][Canto and Nedel 2020][Emani et al. 2024].

Nesse cenário, integrar dados da APS e da vigilância epidemiológica é fundamental para viabilizar o uso dessas bases em pesquisa e inovação tecnológica. Este estudo propõe uma metodologia de preparação, deduplicação e relacionamento entre registros do Prontuário Eletrônico do Cidadão (PEC) e do Sistema de Informação de Agravos de Notificação da tuberculose (SINAN-TB) no município de Manaus. O objetivo é consolidar uma base clínica integrada e rastreável, servindo como alicerce para o processamento de narrativas clínicas e o treinamento de modelos de linguagem no domínio da saúde.

2. Metodologia

A Figura 1 ilustra as etapas de processamento para a identificação de registros pareados presentes simultaneamente em ambas as bases de dados.

2.1. Bases de dados

Foram utilizadas duas bases de dados do Sistema Único de Saúde (SUS) no período de 2019 a 2023: i. PEC (e-SUS APS): Base da APS contendo 7.803.054 registros

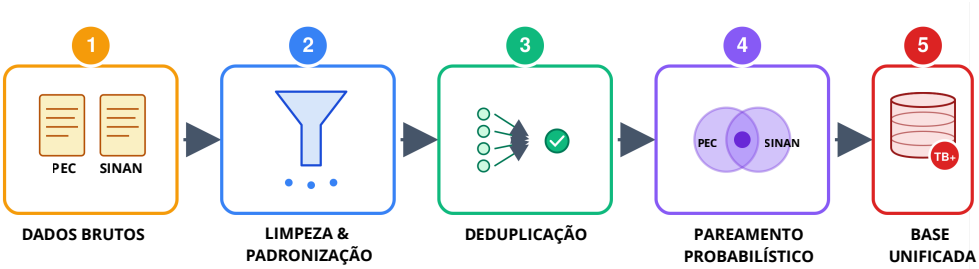

Figura 1. Pipeline do Pareamento Probabilístico para unificação de registros individuais entre PEC e SINAN-TB

de atendimentos em Manaus. Inclui dados de identificação, datas e registros clínicos no padrão SOAP (Subjetivo, Objetivo, Avaliação e Plano). ii. SINAN-TB: Base de notificações compulsórias com 25.186 registros de tuberculose, contendo variáveis demográficas e clínicas. Por restrições éticas (CEP CAAE 86219825.6.00000.0005) e diretrizes de segurança institucional, os dados brutos e os scripts do pipeline são restritos, mas podem ser disponibilizados mediante requisição justificada aos autores. Pela falta de identificadores universais, aplicou-se pareamento probabilístico (PP) por similaridade textual e temporal.

2.2. Pré-processamento e padronização

O pré-processamento buscou a seleção de casos novos de TB pulmonar e a homogeneização das bases para reduzir o ruído na comparação. As etapas incluíram: i. Filtros de qualidade, com a remoção de registros com campos-chave nulos (nome, nome da mãe e data de nascimento) e restrição temporal (data de nascimento $\geq$ 1910). ii. Tratamento de placeholders, com a identificação e exclusão de termos genéricos típicos do sistema (ex: "RN", "IGN", "DESCONHECIDO"). No PEC, exigiu-se conteúdo clínico mínimo em ao menos um campo SOAP. iii. Normalização textual, como conversão para caixa alta, remoção de acentuação, remoção de stopwords (artigos e preposições) e de espaços múltiplos.

2.3. Deduplicação

Cada base foi deduplicada independentemente para garantir que cada indivíduo fosse representado por um único registro antes do cruzamento. Dada a natureza longitudinal do PEC, 93% dos registros eram duplicatas exatas de indivíduos em diferentes atendimentos. A estratégia foi dividida em: i. Determinística: agregação por tríade nominal exata (nome, nome da mãe e data de nascimento), reduzindo a base para 1.521.231 indivíduos únicos (redução de 78,4%). ii. Probabilística: aplicou-se pareamento probabilístico para capturar variações tipográficas (ex: nomes abreviados ou grafias divergentes). A blocagem foi realizada pela data de nascimento completa, decisão tomada pela viabilidade computacional.

A similaridade foi calculada via distância de Levenshtein (limiar $\geq$ 0,60). Este limiar, embora numericamente baixo, é compensado pelo alto grau de concordância estrutural imposto pela blocagem determinística da data. Para resolver cadeias de duplicatas (ex: registros A=B e B=C), o problema foi modelado como a busca de componentes conexos em um grafo de registros, onde pares confirmados representam arestas. Utilizou-se o algoritmo Union-Find (via NetworkX) para agrupar as famílias de duplicatas. Foram identificadas 26.065 famílias, das quais 98,2% eram pares simples. Para cada compo-

nente, elegeu-se como representante o registro cronologicamente mais antigo, garantindo a preservação do identificador original de entrada no sistema.

Já a estratégia para o SINAN-TB, pelo volume reduzido, aplicou-se blocagem mais flexível (sexo e ano de nascimento). Utilizou-se a métrica Jaro-Winkler com penalidades para erros críticos (diferença > 2 caracteres). Os pares foram estratificados em zonas de confiança, com revisão manual da "zona cinzenta"(similaridade entre 0,77 e 0,84) para assegurar a integridade do padrão-ouro.

2.4. Pareamento probabilístico entre PEC e SINAN

O cruzamento entre o PEC (1,5M) e o SINAN-TB (11,9 mil) utilizou blocagem por ano de nascimento. Essa escolha foi estratégica para mitigar erros de preenchimento de dia e mês, permitindo recuperar 1.063 casos (10,7% do total pareado) que seriam perdidos em uma blocagem estrita de data completa.

O processamento dos 245 milhões de pares candidatos foi realizado em lotes paralelizados. A métrica de comparação foi a média da similaridade de Levenshtein entre os nomes do paciente e da mãe. Para casos de múltiplos candidatos no PEC para um mesmo registro do SINAN-TB, selecionou-se o par com maior score. Nesta etapa preliminar da pesquisa, adotou-se um limiar de confiança conservador ($Weight \geq 0,90$.) após inspeção visual da distribuição de scores.

O pipeline foi implementado em Python 3, utilizando Polars para manipulação de dados em memória e as bibliotecas RecordLinkage, Jellyfish e NetworkX. O código segue princípios de pesquisa reprodutível, com estrutura modular e documentação técnica.

3. Resultados parciais

A execução do pipeline sobre as bases do PEC e SINAN-TB permitiu a consolidação de um ecossistema de dados unificado, reduzindo a redundância informacional e integrando trajetórias assistenciais fragmentadas. Os resultados são analisados conforme o fluxo volumétrico de processamento e o PP.

3.1. Processamento e Compressão de dados

O impacto das etapas iniciais de preparação dos dados é detalhado na Figura 2. A transição da Base bruta para a etapa de pós-deduplicação revelou naturezas distintas entre os sistemas: i. PEC (e-SUS APS): Partindo de 7,8 milhões de registros brutos, a etapa de limpeza removeu inconsistências estruturais iniciais, mas foi na etapa da deduplicação que se gerou a maior compressão, consolidando os dados em 1,5 milhão de indivíduos únicos. Este resultado confirma que o PEC opera como um registro por evento clínico (atendimento), exigindo uma engenharia de dados robusta para reconstruir o percurso longitudinal do paciente. ii. SINAN-TB: Diferente do PEC, a base de notificações apresentou baixa duplicação interna, refletindo sua característica de sistema de registro por agravo e não por atendimento rotineiro.

Dos 25,2 mil registros iniciais, a redução para 11,9 mil notificações únicas concentrou-se na etapa de limpeza; esse decréscimo resultou do recorte temporal (2019-2023) e da exclusão da tuberculose extrapulmonar. Por apresentar clínica e fisiopatologia distintas e amostra inadequada, essa decisão metodológica garantiu a fidelidade do pareamento, priorizando a forma com maior densidade na APS.

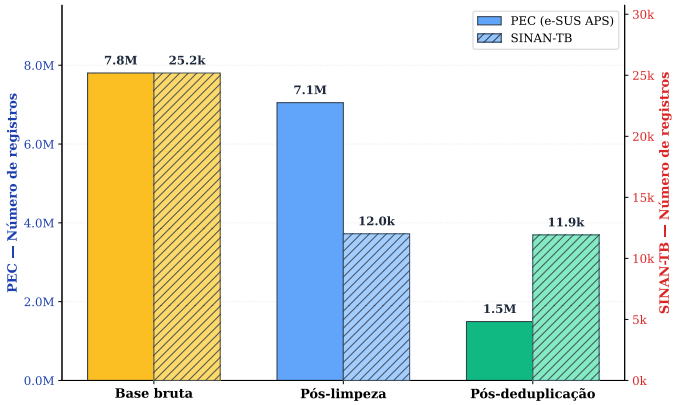

Figura 2. Fluxo volumétrico de registros: dos dados brutos à união de indivíduos únicos após deduplicação.

3.2. Pareamento probabilístico e Base unificada

A integração entre o PEC (1,5M) e o SINAN-TB (11,9 mil) utilizou um PP para superar a ausência de identificadores únicos. Em relação ao desempenho do pareamento, conforme ilustrado na Figura 3, a distribuição dos pesos de similaridade (Weights) demonstrou alta consistência interna: 81,9% dos pares identificados atingiram um limiar de altíssima confiança ($\geq$ 0,95). Na interseção das bases (Figura 4), o processo identificou 8.596 registros pareados, o que representa uma cobertura de 72,1% do universo de notificações do SINAN-TB presentes na atenção primária.

O resultado final é uma base unificada que viabiliza o treinamento supervisionado de LLMs ao fornecer o texto de entrada (narrativas clínicas SOAP do PEC) rotulado com o desfecho confirmatório (dados epidemiológicos do SINAN-TB), permitindo que os modelos aprendam padrões linguísticos de suspeição de tuberculose. Esta integração permite identificar três grupos críticos para essas tarefas de classificação: i. Positivos Confirmados: Registros com diagnóstico no SINAN e acompanhamento no PEC. ii. Negativos/Não Notificados: Indivíduos no PEC sem correspondência no SINAN (úteis para detecção de subnotificação). iii. Perdas de Seguimento: Registros presentes no SINAN sem correspondente no PEC, pois o diagnóstico ou acompanhamento ocorreu em níveis secundário ou terciário, cujos fluxos operam de forma isolada da APS.

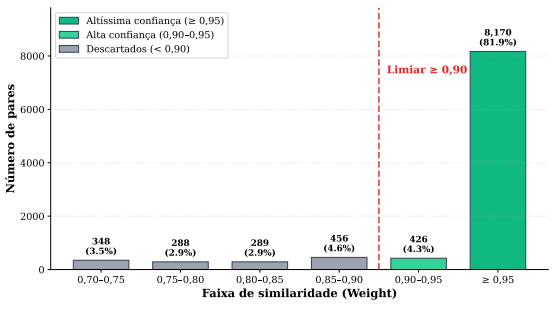

Figura 3. Distribuição de pares por similaridade e limiar ($\geq$ 0,90).

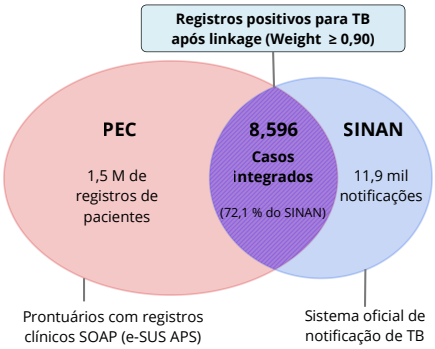

Figura 4. Interseção das bases: representação dos 8.596 casos integrados.

4. Considerações finais

Este estudo estabeleceu uma metodologia para a integração de registros da Atenção Primária e Vigilância Epidemiológica, demonstrando que o rigor no pareamento probabilístico extrai dados robustos de fontes ruidosas. Com um índice de pareamento de 72,1%, a base resultante não é apenas um repositório, mas o alicerce para LLMs voltados ao suporte à decisão clínica e à vigilância oportuna. Essa infraestrutura permite que ferramentas inteligentes realizem a triagem automática de narrativas, auxiliando profissionais na identificação de casos suspeitos em tempo real. Trabalhos futuros focarão na validação da acurácia do pipeline e na aplicação de técnicas de PLN sobre campos SOAP, assegurando fidelidade clínica e mitigação de alucinações em cenários complexos de saúde pública.

5. Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001, do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e da Fundação de Amparo à Pesquisa do Estado do Amazonas (FAPEAM). Vanderson Sampaio é bolsista produtividade CNPq (Chamada 18/2024).

Referências

Canto, V. B. and Nedel, F. (2020). Completeness of tuberculosis records held on the notifiable health conditions information system (sinan) in santa catarina, brazil, 2007-2016. *Epidemiologia e Serviços de Saúde*, 29(3).

Emani, S. et al. (2024). Quantifying gaps in the tuberculosis care cascade in brazil: a mathematical model study using national program data. *PLOS Medicine*, 21.

Hossain, E. et al. (2023). Natural language processing in electronic health records in relation to healthcare decision-making: a systematic review. *Computers in Biology and Medicine*, 155:106649.

Lima, S. et al. (2020). Quality of tuberculosis information systems after record linkage. *Revista Brasileira de Enfermagem*, 73(5).

Meckawy, R. et al. (2022). Effectiveness of early warning systems in the detection of infectious diseases outbreaks: a systematic review. *BMC Public Health*, 22.

Moreira, S. F. et al. (2025). Integração de bases de dados para qualificação da vigilância da tuberculose. In *Saúde coletiva na Amazônia: pesquisa, formação e territorialidades*, page 417. Rede Única, Porto Alegre.

Sacramento, D. S. et al. (2026). Digital solutions for tuberculosis surveillance and control in primary health care: a scoping review. *Epidemiologic Reviews*. Accepted manuscript.

Sampaio, V. S., Lopes, R., Ozahata, M. C., et al. (2023). Thinking out of the box: revisiting health surveillance based on medical records. *Antimicrobial Stewardship & Healthcare Epidemiology*, 3(1):e185.

Seinen, T. et al. (2024). Using structured codes and free-text notes to measure information complementarity in electronic health records: Feasibility and validation study. *Journal of Medical Internet Research*, 27.