Integração e Visualização de Dados de Saúde Materna e Infantil: Uma Arquitetura Computacional para Análise e Suporte à Tomada de Decisão

Ricardo Morsoleto[1], Maria C. Batista[1],Vinícius A. Silva[1], Juliano de S. Caliari[1], Simone Mara F. Miranda[1], Hiran Nonato M. Ferreira[1]

[1]Instituto Federal de Educação, Ciência e Tecnologia do Sul de Minas Gerais (IFSULDEMINAS) - Campus Passos
Passos – MG – Brasil

{ricardo.morsoleto, maria1.batista}@alunos.ifsuldeminas.edu.br
{vinicius.silva, juliano.caliari}@ifsuldeminas.edu.br
sisimaramiranda@gmail.com, hiran.ferreira@ifsuldeminas.edu.br

Abstract. *The increasing volume of data generated by health information systems demands new approaches for analysis and visualization, particularly in the context of public health. This paper proposes the development of a computational architecture to integrate data from three major Brazilian Health Information Systems (HIS): SIM (Sistema de Informação sobre Mortalidade), SINASC (Sistema de Informação sobre Nascidos Vivos), and SIH (Sistema de Informação Hospitalar). The primary objective is to facilitate integrated data analysis and visualization, enabling the extraction of insights to support public policy formulation. Partial results demonstrate the effectiveness of the proposal through the extraction and integration of datasets into a unified database, which allows for temporal, regional, and socio-contextual analyses.*

Resumo. *O crescente volume de dados gerados por sistemas de informação em saúde exige novas abordagens para análise e visualização, especialmente no contexto da saúde pública. Este artigo propõe a criação de uma arquitetura computacional para integrar dados de três importantes Sistemas de Informação em Saúde (SIS) brasileiros: SIM (Sistema de Informação sobre Mortalidade), SINASC (Sistema de Informação sobre Nascidos Vivos) e SIH (Sistema de Informação Hospitalar). O objetivo principal é facilitar a análise e visualização dos dados de forma integrada, permitindo a extração de insights para apoiar a formulação de políticas públicas. Resultados parciais mostram a efetividade da proposta a partir da extração e integração das bases em uma base unificada, a qual permite recortes temporais, regionais e de variados contextos sociais.*

1. Introdução

Com o aumento exponencial da quantidade de dados gerados diariamente em prontuários eletrônicos e Sistemas de Informação em Saúde (SIS), torna-se essencial a adoção de métodos avançados de análise de dados para a extração de insights relevantes [Adeniran et al. 2024]. No Brasil, os SIS desempenham um papel fundamental na formulação de políticas públicas, fornecendo informações essenciais para a tomada de decisões na gestão da saúde [Brasil and da Saúde 2009].

Entretanto, esses dados encontram-se distribuídos em diferentes bases, o que exige processos de integração para que possam ser utilizados de forma eficiente na tomada de decisão [Sousa et al. 2019]. Contudo, os SIS foram originalmente concebidos para operar de maneira independente, sem a disponibilização de bases previamente integradas, tornando a vinculação dos dados um processo manual e desafiador.

Para realizar essa integração, é necessário um pré-processamento dos dados, que inclui etapas como limpeza e padronização, visando minimizar erros e garantir a qualidade das informações [Han et al. 2022]. Além disso, um dos principais obstáculos para essa tarefa é a ausência de atributos primários, como CPF ou nome completo, que facilitariam a correspondência entre registros de diferentes bases.

A integração dos SIS possibilita a utilização dessas informações em ferramentas de visualização de dados, permitindo a análise da trajetória de tratamento do paciente (TTP). A TTP refere-se ao conjunto de consultas, diagnósticos e tratamentos de um paciente ao longo do tempo, sendo essencial para o monitorar os serviços de saúde utilizados [Pinaire et al. 2017]. Um exemplo prático da TTP é a análise da mortalidade infantil, que exige a vinculação entre os sistemas de nascidos vivos e de mortalidade.

Diante desse cenário, este estudo propõe o desenvolvimento de uma arquitetura computacional para facilitar a utilização dos dados dos SIS por meio da visualização de dados. Espera-se que essa abordagem contribua para o aprimoramento do monitoramento da saúde pública e para a maior agilidade na formulação de políticas públicas.

2. Background

No Brasil, os Sistemas de Informação em Saúde (SIS) foram implantados para garantir o planejamento de novas políticas públicas e apoiar a inovação na gestão da saúde [Brasil and da Saúde 2009]. Existem, ao todo, mais de 15 SIS voltados para a área da saúde, que podem ser acessados e consultados por meio do site do Departamento de Informática do SUS (DataSUS)[1]. Dentre eles, o foco deste trabalho é a integração do Sistema de Informações sobre Nascidos Vivos (SINASC), do Sistema de Informações sobre Mortalidade (SIM) e do Sistema de Informações Hospitalares do SUS (SIH/SUS).

O SINASC[2], criado em 1990, tem como principal função coletar informações sobre nascimentos no Brasil. Seus dados são organizados em duas bases: a Declaração de Nascidos Vivos (DN), com registros de 1994 a 2023 e cobertura nacional, e a Declaração de Nascidos Vivos Residentes no Exterior (DNEX), que abrange o período de 2014 a 2023. O sistema apresenta uma cobertura superior a 90%, o que o torna uma fonte confiável para estudos sobre indicadores de saúde [Pedraza 2012]. Entre as informações disponíveis, destacam-se características maternas, como idade, tipo de parto e cor, além de dados do recém-nascido, como peso ao nascer, índice de APGAR no quinto minuto e local de nascimento. Devido à sua abrangência, o SINASC é amplamente utilizado em pesquisas sobre saúde materno-infantil [Guerra et al. 2008, Falavina et al. 2024].

O SIM[3], criado em 1975, tem como finalidade registrar informações sobre óbitos no país. O sistema coleta dados sobre as causas de morte, características do falecido

[1]https://datasus.saude.gov.br/transferencia-de-arquivos/

[2]https://www.gov.br/saude/pt-br/composicao/svsa/sistemas-de-informacao/sinasc

[3]https://www.gov.br/saude/pt-br/composicao/svsa/sistemas-de-informacao/sim

(como idade, sexo e local de residência) e circunstâncias do óbito. Essas informações são essenciais para estudos epidemiológicos e para a construção de indicadores de mortalidade, incluindo a taxa de mortalidade infantil [Brasil and da Saúde 2009]. O SIM também é amplamente utilizado em pesquisas voltadas à análise de causas específicas de óbito [Guedes et al. 2023].

Por sua vez, o SIH/SUS[4] foi implantado na década de 1990 com o objetivo de registrar as Autorizações de Internação Hospitalar (AIH). Esse sistema armazena informações detalhadas sobre internações hospitalares, incluindo diagnósticos, procedimentos realizados, tempo de permanência e custos envolvidos. Além de ser um instrumento essencial para a gestão de recursos hospitalares, o SIH é amplamente utilizado em estudos sobre morbidade hospitalar no Brasil [Domingues et al. 2024]. Pesquisas baseadas nessa base de dados frequentemente abordam a evolução de quadros clínicos e os impactos de doenças na rede hospitalar [Azevedo e Silva et al. 2014, Alves et al. 2024].

Esses sistemas são independentes entre si, cada um com sua própria estrutura de dados. Essa característica impõe desafios à integração das bases por meio do Record Linkage, já que não existe um identificador único entre os bancos, além dos dados serem disponibilizados de forma anonimizada, sem atributos primários, como o número do CPF ou o nome do paciente. Mesmo com esses desafios, foi possível realizar a integração das bases SINASC, com 31.351.323 registros e 67 variáveis, e SIM (apenas mortalidades menores que 1 ano), com 382.777 e 99, gerando um banco com 266.625 registros totais.

3. Abordagem Proposta

A proposta deste trabalho consiste na implementação de uma abordagem computacional para a integração e visualização de dados oriundos dos Sistemas de Informação em Saúde (SIS) brasileiros, especificamente o SIM, o SINASC e o SIH/SUS. O objetivo central é permitir que os dados sejam coletados, integrados e disponibilizados de forma estruturada para facilitar a extração de insights e a formulação de políticas públicas de saúde.

Para a construção da abordagem proposta foi necessária a implementação de atividades específicas de coleta, processamento, armazenamento e análise de dados. O processo pode ser sintetizado em cinco etapas, conforme ilustrado na Figura 1.

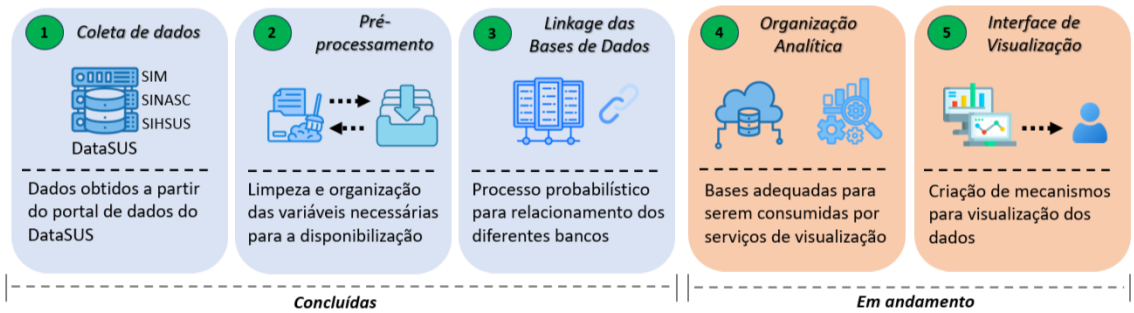

Figura 1. Etapas para construção da abordagem proposta.

A fase de *Coleta de Dados (1)* foi realizada a partir das fontes do Ministério da Saúde, por meio do DataSUS, garantindo a obtenção de informações atualizadas e

[4]http://sihd.datasus.gov.br/principal/index.php

confiáveis. Esse processo envolveu a extração dos dados brutos disponibilizados entre os anos de 2012 e 2023[5] e a implementação de mecanismos para assegurar a atualização contínua desses dados à medida que novas informações forem disponibilizadas.

A fase de *Pré-processamento dos Dados (2)* envolveu uma série de técnicas para garantir a qualidade dos dados antes da integração. Foram aplicados processos de remoção de inconsistências, tratamento de valores ausentes, padronização de formatos e normalização dos dados. Além disso, foram realizadas transformações para unificar as nomenclaturas utilizadas nos diferentes bancos garantindo que os dados estejam prontos para a fase de integração. Essas etapas são fundamentais para evitar redundâncias, eliminar erros e assegurar que os dados estejam organizados de maneira coerente.

A fase de *Linkage das Bases de Dados (3)* foi conduzida utilizando métodos probabilísticos para integrar os registros dos três sistemas, uma vez que eles não compartilham identificadores únicos, como CPF ou nome completo. Para superar essa limitação, foram aplicadas técnicas de linkage baseadas em múltiplos atributos, como nome da mãe, data de nascimento e município de residência [Omitido, 2024]. Esse processo permitiu a vinculação dos registros, possibilitando a criação de um banco de dados integrado que amplia o potencial de análise sobre os dados registrados nos diferentes sistemas.

Na fase de *Organização Analítica (4)*, os dados são segmentados em diferentes recortes analíticos, como faixas temporais, regiões geográficas e categorias específicas de estudo. Algumas análises já foram conduzidas, mas a API para disponibilização estruturada dos dados ainda não foi desenvolvida e segue em fase de implementação.

Por fim, a fase de *Interface de Visualização (5)* contempla a criação de interfaces interativas para exploração dos dados. Essas interfaces poderão ser utilizadas por diferentes tipos de dashboards e sistemas de análise, proporcionando aos pesquisadores e gestores de saúde ferramentas avançadas para a tomada de decisão. É importante destacar que esta fase oferece flexibilidade para ser implementada por qualquer interessado, sem a necessidade de depender exclusivamente de uma solução específica. Os participantes podem optar por utilizar ferramentas já existentes e amplamente utilizadas, como Google Data Studio ou Microsoft Power BI, para criar visualizações interativas e dinâmicas. Alternativamente, os interessados também têm a liberdade de criar sua própria interface de visualização, consumindo os dados disponíveis através da API construída na arquitetura proposta. Isso possibilita um nível de personalização mais avançado, permitindo o desenvolvimento de uma solução sob medida para atender a necessidades específicas.

A implementação desta abordagem visa otimizar a integração dos SIS, reduzindo a complexidade do processo manual atualmente necessário para o acesso consolidado aos dados. Dessa forma, espera-se contribuir para aprimorar o monitoramento da saúde pública e para a formulação de políticas mais eficazes e baseadas em evidências.

4. Resultados Parciais

Os resultados parciais obtidos até o momento refletem um progresso consistente nas etapas iniciais do projeto. A coleta de dados foi realizada a partir das fontes do DataSUS, com informações extraídas dos sistemas SIM, SINASC e SIH/SUS, abrangendo o período

[5]Dados mais recentes ainda não foram disponibilizados até a presente data.

de 2012 a 2023. Esse processo assegurou a obtenção de um conjunto robusto e atualizado de dados, essencial para as análises subsequentes.

Na etapa de pré-processamento, foram adotadas técnicas de limpeza e padronização para garantir a consistência e a qualidade dos dados. A remoção de valores inconsistentes e a unificação de nomenclaturas foram fundamentais para viabilizar a integração entre os três sistemas distintos. O processo de integração de dados foi realizado com o uso de linkage probabilístico, uma abordagem crucial diante da ausência de identificadores exclusivos, como CPF ou nome completo. Por meio de características como nome da mãe e data de nascimento, foi possível realizar a integração das informações, criando uma base única e mais confiável para análise.

O primeiro processo de integração realizado consistiu na união dos dados provenientes do SIM e do SINASC. O SIH/SUS não foi inicialmente usado devido à falta de documentação, o que dificulta o processo de padronização, e a presença de múltiplos registros de uma mesma pessoa em diferentes processos hospitalares. Inicialmente, o SIM foi filtrado para incluir apenas os registros referentes a óbitos de crianças de até 1 ano de idade. A integração foi possível utilizando três colunas comuns a ambos os bancos de dados: PESO (peso ao nascer), CODMUNRES (código de município de residência) e DTNASCIMENTO (data de nascimento do bebê). A união dos registros ocorreu com base na correspondência exata dos valores nessas três colunas, resultando em 327.734 registros pareados. Após a integração, os registros foram manualmente revisados e filtrados com base nas discrepâncias encontradas em outras colunas, como SEXO e RACACOR. Como resultado, obteve-se um total de 266.625 registros finais.

Neste momento, o foco do trabalho está na organização analítica dos dados, que estão sendo agrupados em categorias específicas e faixas temporais relevantes para a pesquisa. Embora algumas análises preliminares já tenham sido realizadas, a construção da API para disponibilização dos dados de forma estruturada ainda está em fase de desenvolvimento. Assim que concluída, a API permitirá que qualquer ambiente computacional adapte sua estrutura para consumir dados e recursos da arquitetura proposta.

5. Considerações Finais

A arquitetura proposta tem mostrado resultados promissores. A conclusão das fases de coleta, pré-processamento e linkage permitiu a criação de uma base de dados integrada de alta qualidade, que já pode ser explorada por meio de análises preliminares.

A fase de organização analítica está em progresso, com a expectativa de que a disponibilização dos dados através da API e a construção de interfaces interativas proporcionem maior acessibilidade e usabilidade dos dados para os gestores de saúde pública. A implementação dessa arquitetura visa otimizar o processo de integração de dados, tornando mais eficiente a análise de informações e a tomada de decisões, especialmente na formulação de políticas públicas baseadas em dados concretos.

Embora a etapa 4 esteja em andamento, os resultados obtidos demonstram a viabilidade da proposta e o potencial impacto que ela pode ter na melhoria da saúde pública, especialmente nas áreas de saúde materno-infantil. A etapa final promete ser um marco importante para a aplicação prática dos dados integrados, permitindo a criação de soluções dinâmicas e personalizadas que atendam às necessidades dos profissionais de saúde.

Agradecimentos

Os autores agradecem o apoio do CNPq (Processo: 445273/2023-2) e do IFSULDEMINAS pelo apoio concedido a este trabalho.

Referências

Adeniran, I. A., Efunniyi, C. P., Osundare, O. S., and Abhulimen, A. O. (2024). Data-driven decision-making in healthcare: Improving patient outcomes through predictive modeling. *Engineering Science & Technology Journal*, 5(8).

Alves, A. F., Pereira, P. H. S., Vidal, T. G. S., Schulz, M. E. B., de Carvalho, L. B., Pinho, M. L. B., and França, J. V. T. (2024). Impacto da doença reumática crônica do coração no brasil: Um estudo pela perspectiva do sus. *Revista Contemporânea*, 4(12):e6891–e6891.

Azevedo e Silva, G., Bustamante-Teixeira, M. T., Aquino, E. M., Tomazelli, J. G., and dos Santos-Silva, I. (2014). Acesso à detecção precoce do câncer de mama no sistema único de saúde: uma análise a partir dos dados do sistema de informações em saúde. *Cadernos de Saúde Pública*, 30:1537–1550.

Brasil and da Saúde, M. (2009). Manual de vigilância do óbito infantil e fetal e do comitê de prevenção do óbito infantil e fetal.

Domingues, R. M. S. M., Meijinhos, L. d. S., Guillen, L. C. T., Dias, M. A. B., Saraceni, V., Pinheiro, R. S., Paiva, N. S., and Coeli, C. M. (2024). Estudo de validação das internações obstétricas no sistema de informações hospitalares do sistema único de saúde para a vigilância da morbidade materna: Brasil, 2021-2022. *Epidemiologia e Serviços de Saúde*, 33:e20231252.

Falavina, L. P., Fujimori, E., and Lentsck, M. H. (2024). Trend of incompleteness of the robson classification variables in the live birth information (sinasc) in the state of paraná, brazil, 2014-2020. *Epidemiologia e Serviços de Saúde*, 33:e2023632.

Guedes, R., Dutra, G. J., Machado, C., and Palma, M. A. (2023). Avaliação dos dados de mortes por covid-19 nas bases dos cartórios do rc-arpen, sivep-gripe e sim no brasil em 2020. *Cadernos de Saúde Pública*, 39(3):e00077222.

Guerra, F. A. R., Llerena Jr, J. C., Gama, S. G. N. d., Cunha, C. B. d., and Theme Filha, M. M. (2008). Defeitos congênitos no município do rio de janeiro, brasil: uma avaliação através do sinasc (2000-2004). *Cadernos de Saúde Pública*, 24:140–149.

Han, S., Shen, D., Nie, T., Kou, Y., and Yu, G. (2022). An enhanced privacy-preserving record linkage approach for multiple databases. *Cluster Computing*, 25(5):3641–3652.

Pedraza, D. F. (2012). Qualidade do sistema de informações sobre nascidos vivos (sinasc): análise crítica da literatura. *Ciência & Saúde Coletiva*, 17:2729–2737.

Pinaire, J., Azé, J., Bringay, S., and Landais, P. (2017). Patient healthcare trajectory. an essential monitoring tool: a systematic review. *Health information science and systems*, 5:1–18.

Sousa, M. J., Pesqueira, A. M., Lemos, C., Sousa, M., and Rocha, Á. (2019). Decision-making based on big data analytics for people management in healthcare organizations. *Journal of medical systems*, 43:1–10.