

Resumo

O Fundo de Educação Sanitária e Imunização em Massa contra Doenças Transmissíveis (Fesima) é o órgão da Secretaria de Estado da Saúde responsável por apoiar e financiar projetos que promovam e protejam a saúde da população no estado de São Paulo. O presente Resumo cumpre o objetivo de conferir transparência e divulgar os projetos em andamento. Informações sobre o Fesima, bem como a forma de submissão de projetos, estão disponíveis em www.ccd.saude.sp.gov.br

Integração das Bases de Dados HIV + e AIDS-SP

Integration of HIV + and AIDS-SP Databases

Centro de Referência e Treinamento em DST/Aids. Coordenadoria de Controle de Doenças. Secretaria de Estado da Saúde. São Paulo, Brasil.

OBJETIVO: Integrar bases de dados (DB) de: casos de HIV⁺ e AIDS, exames laboratoriais, dispensação de antirretrovirais e óbitos, para subsidiar a atuação técnica do Centro de Referência em DST/AIDS – (CRT-DST/AIDS). **METODO E MATERIAL: ESCOPO E INFRAESTRUTURA:** Propôs-se arquitetura tecnológica de *DATAWAREHOUSE* (DW), gerenciador de DB *PostgreSQL* e, *Pentaho Data Integration- PDI* para ETL (*Extract, Transform e Load*). Solução de *Dataviews*: Tableau ou Power BI. *Hardware: storage* de médio porte, infraestrutura de rede e *conectividade* existente. **EVENTOS:** residentes no Estado, período 1980 -2020. **FONTES:** uso de quatro Sistemas de Informação em Saúde (SIS) 1) **SINAN – Agravos e Doenças de Notificação:** casos de AIDS (campo ID_AGRAVO, código B24 (3C-CID10), casos HIV⁺ (código B24, código 901, campo CRITERIO. 2) **SIM – Mortalidade:** óbito com Causa Básica de Morte de *HIV⁺ E AIDS (B20 a B24 CID10)*. 3) **SISCEL -Exames Laboratoriais:** cadastro e resultados de contagem de linfócitos TCD4⁺/TCD8⁺ e carga viral HIV. 4) **SICLOM - Medicamentos:** cadastro e dispensação de antirretrovirais. **ESTRATÉGIAS OPERACIONAIS:** *Volumetria e Normalização: volumetria* (mensuração de linhas e colunas) e normalização (*tipo, tamanho, caixa, caracteres, encoding*). 2) *Transformação e Aprimoramento: tratamento de datas, remoção de termos* associados a nomes próprios através de expressões regulares (REGEX), *inclusão de novas colunas e, fonetização* para comparação de registros com *Phyton* e bibliotecas *Fonetipy* e *Metaphone*. *De-duplicação:* identificação/remoção de duplicatas de mesmo registro com blocagem (nome, mãe, sexo e DN) e métodos de funções de distância entre *strings*, métodos *determinístico* e *probabilístico* (*Levenstein e Winkler*) com corte de 90% pois não há campo/variável comum nas DB. Três condições: *identificação, seleção e aprimoramento* do registro para finalização. 3) *Pareamento (linkage), Auditoria e Validação:* modelagem *Star Schema* centrada no indivíduo (paciente) detentor do evento. Método *determinístico* e *probabilístico* (*Levenstein e Winkle*), corte de 90% para os *matches* (pares). *Auditoria e validação* dos *matches* se deram por dupla. Formação da DB integrada HIV⁺ e AIDS, com 1º *merge* de *casos/*

óbitos e na sequência exames/medicamentos com finalização da integração. Usou-se *Python* com as bibliotecas *Numpy*, *Pandas* e *Record Linkage*, e *Programa R FRAMEWORK*: criação de interface amigável, ágil e de baixa complexidade para interação de correção e validação na etapa ETL. Usou-se o *Streamlit (open source)* na solução. **FINANCIAMENTO**: montante de R\$ 97.320,00 originários do FESIMA/CCD/SES-SP de outubro/2019 a dezembro/2020. **RESULTADOS**: O volume apurado no conjunto alcançou 456.029 exames laboratoriais e dispensação de antirretrovirais. Os óbitos somaram 67852 e casos 417.762. O SINAN resultou em 170 colunas, SIM em 97 colunas, SISCEL e SICLOM em 140 colunas, todos com as respectivas variáveis na totalidade e, no centro o *indivíduo*. Os resultados representaram esforço técnico da vigilância epidemiológica do CRT DST/AIDS com foco em estudos preditivos e análise de sobrevivência para HIV⁺ e AIDS. **CONSIDERAÇÕES FINAIS**: Dentre os desafios e demandas gerados pelos resultados estão o desenvolvimento da solução de *Dataviews* com interface interativa pela plataforma web com *delivery self service* ao usuário final e a disseminação de informações de boa qualidade e oportuna aos serviços de saúde, pesquisadores e sociedade sobre HIV⁺ e AIDS.

REFERÊNCIAS

1. YAOHAO P, MATION LF - O Desafio do pareamento de grandes bases de dados: Mapeamento de métodos de record linkage probabilístico e diagnóstico de sua viabilidade empírica - Instituto de Pesquisa Econômica Aplicada – IPEA, 2018. Texto para discussão/Instituto de Pesquisa Econômica Aplicada. Brasília: Rio de Janeiro: Ipea, 1990.
2. BRASIL. MINISTÉRIO DA SAÚDE. Secretaria de Vigilância em Saúde. Departamento de Análise de Situação de Saúde. Manual de Instruções para o preenchimento da Declaração de Óbito/ Ministério da Saúde, Secretaria de Vigilância em Saúde, Departamento de Análise de Situação de Saúde. – Brasília: Ministério da Saúde, 2011. 54 p.: il. (Série A. Normas e Manuais Técnicos)
3. BRASIL. MINISTÉRIO DA SAÚDE. Secretaria de Vigilância em Saúde. Departamento de Vigilância Epidemiológica. Sistema de Informação de Agravos de Notificação – Sinan: normas e rotinas/Ministério da Saúde, Secretaria de Vigilância em Saúde, Departamento de Vigilância Epidemiológica. – 2. ed. – Brasília: Editora do Ministério da Saúde, 2007. 68 p.: il.– (Série A. Normas e Manuais Técnicos)
4. DUARTE DAP, CORREA CSL, FAYERVA, NOGUEIRA MC, BUSTAMANTE MT - Avaliação de método para classificação automatizada de pares em relacionamentos probabilísticos de bancos de dados. Cad. Saúde Pública 2019; 35(11):e00066419.p1-10.
5. QUEIROZ OV, GUERRA JÚNIOR AA, MACHADO CJ, ANDRADE EIG, MEIRA JÚNIOR W, ACURCIO FA,
6. SANTOS FILHO W, CHERCHIGLIA ML. Relacionamento de registros de

grandes bases de dados: estimativa de parâmetros e validação dos resultados, aplicados ao relacionamento dos registros das autorizações de procedimentos

ambulatoriais de alta complexidade com os registros de sistema de informações hospitalares. Cad. Saúde Colet., 2010, Rio de Janeiro, 18 (2): 298-308.