PROADI - Hospital Israelita Albert Einstein

Versão preliminar

PROADI - Hospital Israelita Albert Einstein

Qualidade de dados
HÓRUS (vinculaSUS Alagoas)

Histórico de revisões

Data	Versão	Descrição	Autor	Responsável
22/06/2020	1.0	Versão preliminar	Leandro Furlam, Elias Ribeiro	Alexandre Rodrigues

Sumário

1 Qualidade de dados

O processo de análise de qualidade de dados está focado na avaliação de conjuntos de dados e na aplicação de ações corretivas, para garantir que estes estejam adequados aos propósitos para os quais foram originalmente destinados (1). Dessa forma, a qualidade de dados está diretamente relacionada a confiabilidade dos dados de entrada. Considerando que os dados têm níveis inadequados de qualidade, é provável que ocorram erros, que podem se propagar acidentalmente e inconscientemente por todo o fluxo da informação, prejudicando a eficiência do sistema. Formas regulares de avaliar a qualidade de dados com modelos clássicos geralmente se destinam a detectar e corrigir erros em fontes conhecidas com base em um conjunto limitado de regras. No ambiente de *Big Data*, a quantidade de regras pode ser enorme e o custo da aplicação para correção de erros pode não ser viável e nem apropriado (*e.g.* o enorme volume de dados ou a volatilidade dos dados de *streaming*). Isso ocorre principalmente porque o *Big Data* não é apenas sobre dados, mas também sobre uma pilha conceitual e tecnológica completa, incluindo dados brutos e processados, armazenamento, formas de gerenciar dados, processamento e análise (1).

Uma dimensão de qualidade de dados é um termo descritor de um recurso de dados, o qual pode ser medido ou avaliado de acordo com padrões definidos, a fim de determinar a qualidade de um conjunto de dados (2). Geralmente, dados só têm valor quando dão suporte a um processo ou a uma tomada de decisão. Em consequência, as regras de qualidade de dados definidas devem levar em consideração o valor que os dados podem fornecer para o sistema.

Neste relatório, seis dimensões de qualidade de dados são analisadas: completude, conformidade, acurácia, consistência, unicidade e temporalidade (2).

Completude caracteriza a taxa de preenchimento das variáveis. Para cada variável é calculado o percentual de entradas com informação não nulas, respeitando, quando houver, sua dependência com outras variáveis.

Conformidade detecta concordância nos valores digitados nos campos das variáveis, avaliando se os valores de entrada não nulos estão em conformidade com os padrões descritos pelo dicionário de dados. Para cada variável estudada é calculado o percentual de entradas em conformidade com o padrão adotado.

Acurácia visa detectar se informação registrada reflete o evento ou objeto descrito, isto é, verificar se o dado cadastrado está em concordância com o evento observado. Devido ao processo de anonimização dos dados, a análise de acurácia se restringe a verificar a possibilidade das informações registradas. Note que acurácia e conformidade são dimensões distintas, pois enquanto conformidade

avalia o padrão do dado, acurácia avalia a razoabilidade dos dados. Para cada variável estudada é calculado o percentual de entradas com informações acuradas.

Consistência constitui de testes envolvendo duas ou mais variáveis visando detectar inconsistências entre dados de um mesmo registro. Para cada teste considerado é calculado os percentuais de aprovação e falha.

Unicidade objetiva mensurar o grau de duplicidade nos dados, realizando a busca por meio de identificadores dos pacientes.

Temporalidade objetiva efetuar medidas estatísticas nos intervalos de tempos entre eventos, *e.g.* Nascimento de um recém-nascido e inclusão desse registro no sistema. O principal interesse é verificar se o dado é disponibilizado prontamente.

Neste relatório a Seção 2 retrata a base de dados, a Seção 3 define a metodologia, a Seção 4 apresenta os respectivos resultados e a Seção 5 estende a análise com considerações finais.

2 Base de dados

Para qualificar a gestão da Assistência Farmacêutica nas três esferas do SUS, e contribuir para a ampliação do acesso aos medicamentos e da atenção à saúde prestada à população, o Departamento de Assistência Farmacêutica e Insumos Estratégicos do Ministério da Saúde (DAF/SCTIE/MS) apresenta o **HÓRUS - Sistema Nacional de Gestão da Assistência Farmacêutica**. Esse sistema foi inicialmente desenvolvido por meio da parceria estabelecida em 2009 entre DAF/SCTIE, a Secretaria Municipal de Saúde de Recife (SMS/PE), a empresa Pública de Informática de Recife (Emprel), o Departamento de Informática do SUS (DATASUS/SE), o Conselho Nacional de Secretários de Saúde (CONASS) e o Conselho Nacional de Secretarias Municipais de Saúde (CONASEMS) (3).

- **Módulo Básico:** Este módulo permite executar as ações de gestão dos medicamentos do Componente Básico da Assistência Farmacêutica, por meio da realização de movimentações como entradas, distribuições e dispensações. Permite também acompanhar essas ações através da emissão de relatórios contendo informações gerenciais, que subsidiam o planejamento e desenvolvimento das ações de Assistência Farmacêutica na Atenção Básica, disponibilizando, desta forma, informações técnicas necessárias para a qualificação dos serviços e gestão do cuidado. O sistema HÓRUS, em seu módulo básico, atende diversos tipos de serviços que gerenciam medicamentos e insumos. A equipe do HÓRUS no Ministério da Saúde trabalha com o objetivo de acompanhar a implantação do sistema nos municípios, monitorar sua utilização e dar o suporte necessário aos usuários do sistema.

- **Módulo Especializado:** Após a experiência no desenvolvimento e implantação do Hórus-Básico, o Ministério da Saúde, por meio de parceria entre o Departamento de Assistência Farmacêutica e Insumos Estratégicos (DAF/SCTIE/MS) e o Departamento de Informática do SUS (DATASUS/SE/MS), desenvolveu o módulo Especializado do Hórus para gestão do Componente Especializado da Assistência Farmacêutica (CEAF), denominado Hórus-Especializado. O Hórus-Especializado foi concebido para qualificar a gestão do CEAF, possibilitando a realização eletrônica de todas as etapas envolvidas na execução deste Componente.
- **Módulo Estratégico:** O Módulo Estratégico faz parte do Hórus Básico/Estratégico e permite executar e acompanhar as ações de gestão dos medicamentos do Componente Estratégico da Assistência Farmacêutica (CESAF). Nesse sentido, a equipe do Hórus, no Ministério da Saúde, trabalha com o objetivo de acompanhar a implantação do sistema, monitorar sua utilização e dar suporte aos usuários. Para qualificar e ampliar o acesso da população aos medicamentos do Componente Estratégico da Assistência Farmacêutica, o Departamento de Assistência Farmacêutica e Insumos Estratégicos (DAF), do Ministério da Saúde, disponibiliza o Hórus – Sistema Nacional de Gestão da Assistência Farmacêutica. O módulo estratégico, parte integrante do Hórus Básico/Estratégico, permite executar e/ou acompanhar as ações de gestão dos medicamentos dos componentes Estratégico e Básico da Assistência Farmacêutica. Nesse sentido, as equipes do DAF trabalham com o objetivo de acompanhar a implantação do sistema, monitorar a utilização e dar suporte aos usuários de estados e municípios.
- **Módulo Indígena:** Por meio de uma parceria entre a Secretaria de Ciência e Tecnologia e Insumos Estratégicos (SCTIE), a Secretaria Especial de Saúde Indígena (Sesai) e o Departamento de Informática do SUS (DataSUS), foi desenvolvido o módulo do Hórus para a gestão da Assistência Farmacêutica no Subsistema de Atenção à Saúde Indígena (SasiSUS). O Hórus será implantado nos Distritos Sanitários Especiais Indígenas (DSEIs), Polos Base, Casas de Saúde Indígena (Casai) e demais unidades de distribuição e dispensação de medicamentos do SasiSUS. Com a informatização será possível registrar as entradas, saídas e fluxo de produtos de medicamentos na rede de saúde indígena contribuindo para o planejamento, monitoramento, avaliação e execução das ações da Assistência Farmacêutica, com vistas à ampliação do acesso da população indígena aos medicamentos essenciais.

2.1 Informações gerais

Base de dados	HÓRUS (vinculaSUS)
Fonte	<s3://vinculasus-al>
Data de obtenção dos dados	19 de junho de 2020
Período	jan/2011 a out/2019
Região geográfica	Alagoas
Volume	1,8 GB
Número máximo de variáveis	34
Número de registros	6.827.601

3 Métodos

A análise dos dados constitui-se de um esquema cíclico, iniciando no mapeamento da documentação e do comportamento dos dados, através da observação de trechos das bases. Em seguida, são definidas as variáveis de teste. Após, ocorre a obtenção e avaliação dos resultados obtidos, recorrendo, e se necessário retificando, conclusões obtidas nos passos anteriores. O maneio dos dados ocorreu através dos serviços **Amazon Athena** e **Amazon S3**, assim como testes e análises se deu utilizando linguagem R. Os *scripts* utilizados estão disponíveis no repositório de qualidade de dados no **GitHub**.

Enfatiza-se que esses dados podem sofrer alterações, caso ocorram atualizações.

O dicionário de dado utilizado, não oficial, foi inferido das descrições das variáveis contidas nos relatórios de integração disponíveis no sistema *pgadmin* onde o banco de dados foi disponibilizado. O dicionário utilizado é apresentado neste relatório como anexo. No que tange os testes de unicidade, procurou-se analisar apenas as informações individuais dos pacientes.

O Cômputo dos resultados numéricos ocorre de modo cascata, isto é, os registros submetidos ao teste de conformidade devem ser não nulos, os registros submetidos ao teste de acurácia devem estar conformes, os registros submetidos aos testes de consistência devem estar acurados, e quando não for possível, conformes e o mesmo se aplica aos registros submetidos aos testes de unicidade. Em prosseguimento, os resultados numéricos são avaliados nas dimensões analisadas calculando-se a média ponderada dos testes realizados, utilizando como peso o total de registros por variável. Para a consistência, é realizado um ajuste em que todas as variáveis testadas devem existir simultaneamente. Objetivando avaliar a

base de dados, o conjunto de resultados representando cada dimensão foi classificada como excelente (> 90%), ótimo (75% - 89,9%), regular (50% - 74,9%) ou ruim (< 49,9%), baseado nos relatórios do livro *Saúde Brasil*, organizado pela Secretaria de Vigilância em Saúde (4). Em decorrência do método cascata utilizado, é realizado o produto dos resultados obtidos, na Seção 5, caracterizando a qualidade da base de dados como um todo, que também pode ser classificada considerando as classes definidas em *Saúde Brasil* (4).

4 Resultados

Esta Seção apresenta e avalia os resultados dos testes aplicados. As considerações são apresentadas nas subseções a seguir, uma para cada dimensão.

Descrições das variáveis e das bases as quais pertencem são apresentadas no Apêndice A. Resultados numéricos dos testes de completude, conformidade e acurácia são exibidos no Apêndice B. O Apêndice C expõe uma tabela contendo valores atípicos[1] de variáveis quantitativas, isto é, registros numéricos que apresentam grande afastamento em relação aos demais, dentro do universo de uma única variável. Descrições dos testes de inconsistência realizados, bem como seus respectivos resultados numéricos estão descritos no Apêndice D. Os resultados agregados por ano foram obtidos em relação à variável *dt_atendimento*.

4.1 Completude

Nesta dimensão são detectados valores faltantes através da busca pelas constantes representando valores ausentes. Nesse sentido, considerou-se como incompletos os registros contendo os valores *NA*, constante lógica que indica valor ausente em linguagem R, e *NULL*, que representa objetos nulos.

No que tange a distribuição temporal dos resultados (Gráfico 1), houve constância, com todos próximos a 89%.

[1] <https://www.rdocumentation.org/packages/grDevices/versions/3.6.2/topics/boxplot.stats>

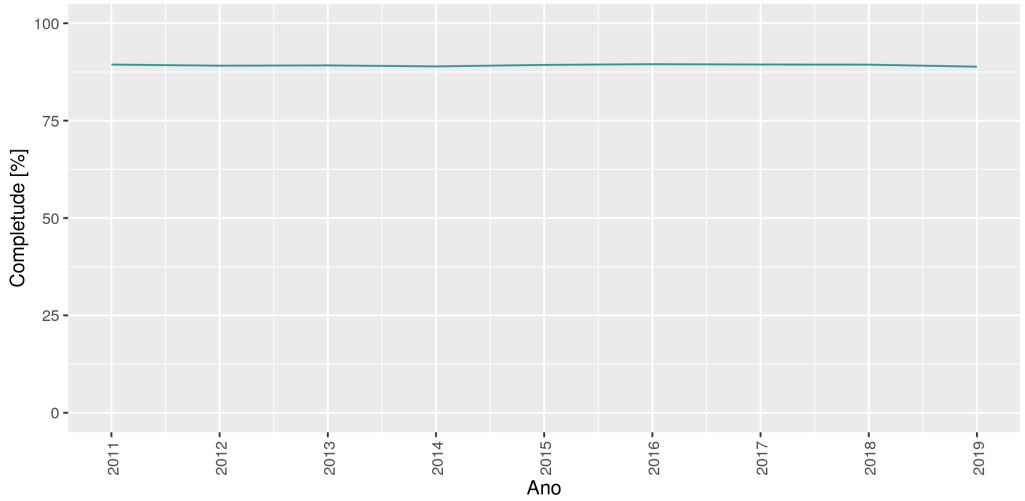

Gráfico 1: distribuição temporal da completude.

No geral, os resultados de completude das variáveis estão distribuídas pelas categorias definidas na Seção 3 segundo o Gráfico 2. O resultado percentual por variável está descrito no Apêndice B. O cômputo da média ponderada dos resultados obtidos é de **89,22%**, ou seja, a **completude é ótimo**.

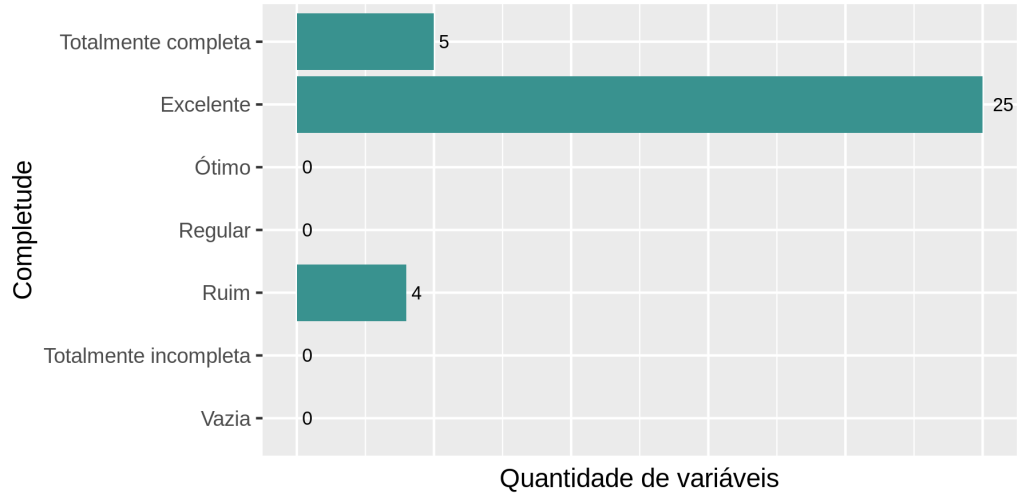

Gráfico 2: distribuição dos resultados de completude.

4.2 Conformidade

Verificou-se se os dados apresentam os padrões descritos no dicionário de dados adotado como referência (Apêndice A) a respeito da quantidade de caracteres. O resultado percentual por variável está descrito no Apêndice B, onde nota-se que todas as variáveis apresentaram conformidades totais em seus registros, **100,00%**, ou seja, a **conformidade é excelente** para a base analisada.

Ressalta-se que, tratando-se de uma base vinculada, as variáveis contidas pos-

suem nomenclaturas diferentes das bases que deram origem à vinculação. Em consequência, não foi possível obter os microdados equivalentes contendo, quando existente, os valores válidos de domínio para cada uma das variáveis. Nesse sentido, a Tabela 1 apresenta a quantidade de registros distintos e no máximo 5 registros mais frequentes por variável, separados por vírgulas.

Variável	**Qtd. distintos**	**Registros mais frequentes**
co_concentracao_medic	17	1, 15, 3, 14, 7
co_formula_farmaceutica	98	1, 5, 4, 8, 14
co_municipio_ibge	102	270430, 270410, 270915, 270470, 270030
co_principio_ativo_medicam	968	63, 34, 68, 981, 11
co_produto	2804	130, 1801, 17, 211, 847
co_seq_paciente	748230	739695, 40540, 1379851, 1471113, 1413625
co_tipo_produto	2	1, 101
co_unidade_consumo	25	4, 11, 1, 9, 10
co_unidade_medida_item	52	21, 31, 12, 66, 32
co_unidade_medida_produto	49	21, 31, 12, 66, 32
co_volume_medicamento	11	1, 6, 181, 5, 241
ds_concentracao_medicamento	17	MG, UI/ML, MG/ML, MCG, UI
ds_forma_farmaceutica	98	COMPRIMIDO, CÁPSULA , SOLUÇÃO ORAL, SUSPENSÃO ORAL , SUSPENSÃO INJETÁVEL
ds_principio_ativo_medicamento	966	HIDROCLOROTIAZIDA , CLONAZEPAM, IBUPROFENO , LOSARTANA POTÁSSICA, AMITRIPTILINA, CLORIDRATO
ds_produto	2802	HIDROCLOROTIAZIDA 25MG COMPRIMIDO, LOSARTANA POTÁSSICA 50 MG COMPRIMIDO, AMITRIPTILINA, CLORIDRATO 25 MG COMPRIMIDO ELENCO ESTADUAL , OMEPRAZOL 20 MG CÁPSULA , INSULINA HUMANA NPH 100 UI/ML SUSPENSÃO INJETÁVEL 10 ML ELENCO ESTADUAL
ds_tipo_produto	2	MEDICAMENTO, PRODUTO PARA SAÚDE
dt_atendimento	3034	2019-09-12, 2019-08-19, 2019-09-26, 2019-06-19, 2019-08-08
dt_nascimento	36482	1958-05-10, 1944-06-10, 1945-01-10, 1899-12-30, 1956-04-20
dt_receita	3859	2019-09-19, 2019-09-17, 2019-08-08, 2019-09-03, 2019-09-12

Variável	Qtd. distintos	Registros mais frequentes
id_paciente	662574	22300122200, 23227523300, 2237861900, 26991551400, 11758633800
nu_concentracao_medic	719	20 MG, 25 MG, 5 MG, 500 MG, 100 MG
nu_dias_dispensar	259	30, 60, 1, 5, 7
nu_fator_correcao	58	1, 100, 1000, 500, 120
nu_produto	2801	BR0267674U0042, BR0268856U0042, BR0267512U0042, BR0267712U0041, BR0271157U0063
nu_volume_medicamento	114	10, 100, 20, 30, 60
qt_dose	755	1, 2, 10, 5, 3
qt_duracao_tratam_dia	300	30, 60, 1, 5, 7
qt_posologia	131	1, 2, 3, 4, 6
qt_solicitado	898	30, 60, 1, 90, 2
sg_tipo_produto	2	M, I
st_rename	2	S, N
tp_produto	4	B, O, S, E
tp_sexo	3	F, M, I
vl_item_dispensacao	19433	0.03, 0.05, 0.04, 0.041, 0.02

Tabela 1: quantidade de registros distintos e registros mais frequentes por variável.

4.3 Acurácia

Inicialmente a métrica de acurácia foi aplicada a dois tipos de registros: datas, ao verificar se o dado configura-se uma data válida e condizente ao período representado pela base de dados; nomes e códigos de municípios, ao verificar se estão contidos na tabela de códigos de municípios e estados do IBGE[2]. Após esta análise de datas e municípios é realizada investigação acerca do preenchimento das variáveis com o objetivo de detectar a presença de preenchimentos sem informações relevantes. Em seguida, são verificados os registros representando informações numéricas, a respeito do sinal (*e.g.* número de filhos deve ser positivo) e do conjunto ao qual pertence (*e.g.* número de filhos deve ser um número inteiro).

Acerca dos valores atípicos descritos no Apêndice C, houve ocorrência em todas as variáveis numéricas analisadas, como por exemplo, posologia de 500 dias e tratamentos de 999 dias. Ainda nesse cenário, cita-se a variável representando a

[2]<https://www.ibge.gov.br/explica/codigos-dos-municipios.php>

data de nascimento do paciente *dt_nascimento*. Nesta, há 1327 registros cujo ano data antes de *1900*, sendo nestes 10 datados de 1850 e atendidos a partir de 2016, ou seja, 10 pacientes atendidos com no mínimo 166 anos.

Acerca das duas falhas encontradas, há 930 ocorrências de ␣ (um espaço em branco) na variável *nu_volume_medicamento* distribuídas a partir de 2013 e 39 ocorrências de *I* (comumente representando *ignorado*) em *tp_sexo* distribuídas a partir de 2014. Estas falhas pouco interferem no resultado final, visto que possuem uma magnitude muito menor em relação ao total de registros.

Mesmo que apresentando erros, estes tampouco interferiram no resultado geral. Nesse contexto, a acuracidade das variáveis estão distribuídas pelas categorias definidas na Seção 3 segundo o Gráfico 3. O resultado percentual por variável está descrito no Apêndice B. O cômputo da média ponderada, considerando todos os testes realizados, é **100,00%**, ou seja, a **acurácia é excelente** para a base analisada.

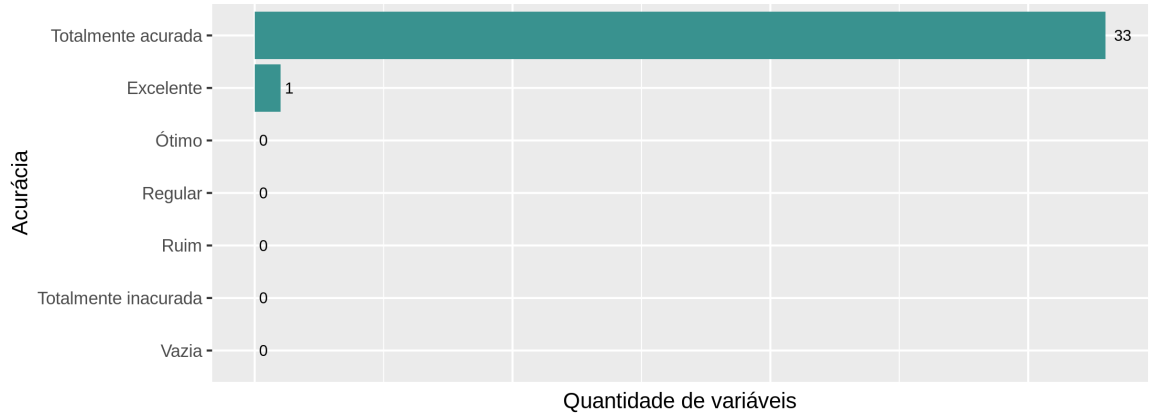

Gráfico 3: distribuição dos resultados de acurácia.

4.4 Consistência

Os resultados descritos na Tabela 2 são de testes aplicados a um mesmo registro, ou seja, mesma linha do conjunto de dados. Estes testes detectam principalmente problemas na entrada de dados envolvendo condições específicas de inconsistências. A descrição dos testes se encontram no Apêndice D. Os resultados dos testes de inconsistência também se encontram detalhados no Apêndice D, destacando a quantidade de inconsistências por teste para cada ano.

Descrição	**Falhas [partes por mil]**
dt_receita >dt_atendimento	1,406
dt_nascimento >dt_receita	0,118
dt_nascimento >dt_atendimento	0,034
ds_tipo_produto == MEDICAMENTO & sg_tipo_produto != M	0,000
ds_tipo_produto == PRODUTO PARA SAÚDE & sg_tipo_produto != I	0,000

Tabela 2: resultados de consistência.

Sobre a distribuição temporal dos resultados de consistência, têm-se o Gráfico 4, onde nota-se que o ano com maior quantidade de registros inconsistentes foi o ano de 2011, com 3,077 partes por mil. Percebe-se que os conforme o passar dos anos as inconsistências diminuem uniformemente.

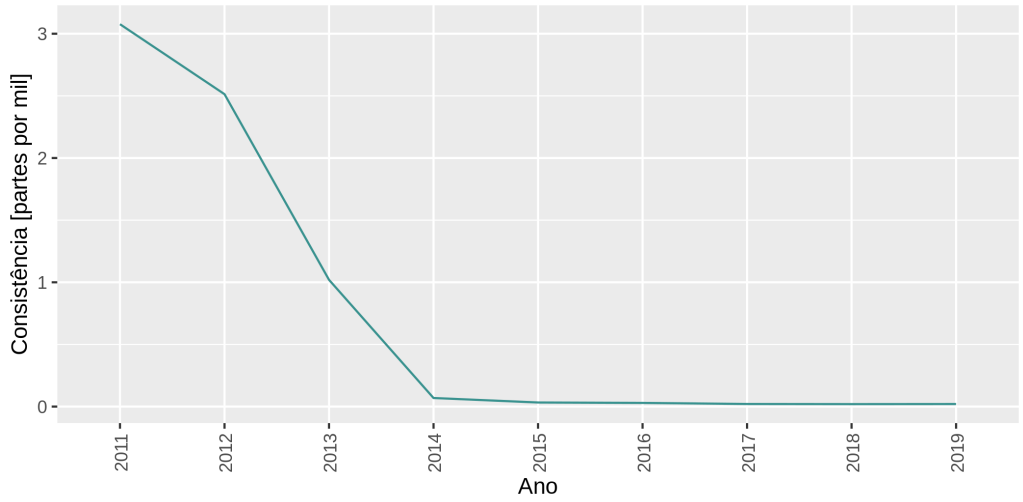

Gráfico 4: distribuição temporal de inconsistências.

No geral, a média ponderada dos resultados dos testes de consistência é **99,97%**, ou seja, a **consistência é excelente** para a base analisada.

4.5 Unicidade

Nesta dimensão é calculado o grau de duplicidade dos dados, buscando diferenças por meio dos identificadores dos pacientes. O identificador, *id_paciente*, foi relacionado a dois tipos de registros: data de nascimento e sexo, representados pelas variáveis *dt_nascimento* e *tp_sexo*, respectivamente. Os registros com respostas *I* (*ignorado*) foram excluídas do cálculo da dimensão pois provocam ambiguidade, gerando sempre duplicidade. A quantidade de registros da variável *tp_sexo* que representam a resposta *I* (*ignorado*) é 0,00057%.

Ao todo, foram encontrados 662.574 identificadores de pacientes distintos e 523.544 identificadores com mais de um registro, sendo que do mesmo, em 0,70% e 0,15%, houve divergência dos registros de sexo e data de nascimento, respectivamente. No geral, aproximadamente 0,66% dos identificadores constam pelo menos um dos erros.

Sobre a distribuição temporal dos resultados gerais dos testes de unicidade, que apresentam os registros no qual constam pelo menos um dos erros, têm-se o Gráfico 5, onde nota-se que todos se aproximam de 100%, sendo a menor unicidade registrada no ano 2010 com 99,50% e a maior no ano de 2019 com 99,85%.

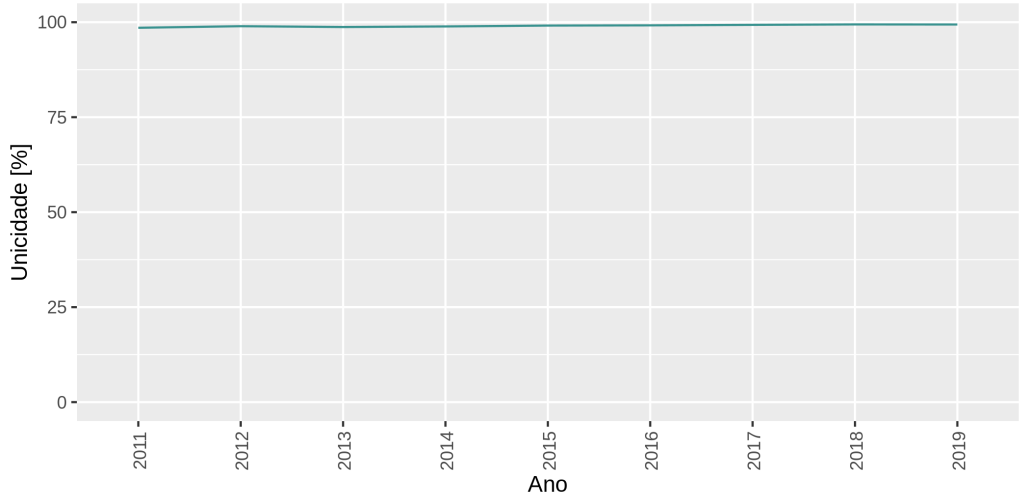

Gráfico 5: distribuição temporal dos resultados de unicidade.

A Tabela 3 apresenta a distribuição de frequência de indicadores dos paciente considerando todos os registros da base de dados. Por exemplo, a quantidade associada a categoria *ID = 1* representa o percentual de pacientes com apenas 1 registro no sistema, enquanto a quantidade associada a a categoria *ID = 2* representa o percentual de pacientes com 2 registros. Mesma interpretação pode ser realizada para as demais categorias. Note que aproximadamente 79,02% dos pacientes tem pelo menos mais de um registro.

Categoria	**Total de identificadores [%]**
ID = 1	20,98
ID = 2	16,33
ID = 3	10,33
ID = 4	7,96
5 <= ID <10	19,32

Categoria	Total de identificadores [%]
10 <= ID <50	21,40
50 <= ID <100	2,67
ID >= 100	1,01

Tabela 3: frequência de identificadores.

No geral, a média ponderada dos resultados dos testes de unicidade é **99,57%**, ou seja, a **unicidade é excelente** para a base analisada.

4.6 Temporalidade

Para mensurar esta dimensão é calculada a quantidade de dias entre duas variáveis representando datas que estejam conformes, acuradas e consistentes. Os resultados são apresentados na Tabela 4. O primeiro e único teste, *T1*, refere-se ao período emissão da receita a data de atendimento do paciente.

Teste	Variável inicial	Variável final	Mediana	Máx.	Min.
T1	*dt_receita*	*dt_atendimento*	0	37000	0

Tabela 4: resultados de temporalidade.

Buscando avaliar os resultados em uma escala temporal, o Gráfico 6 apresenta os resultados das medianas em relação ao ano representado. Nota-se que houve um resultado uniforme da mediana em relação aos anos, mesmo resultado que foi a mediana geral.

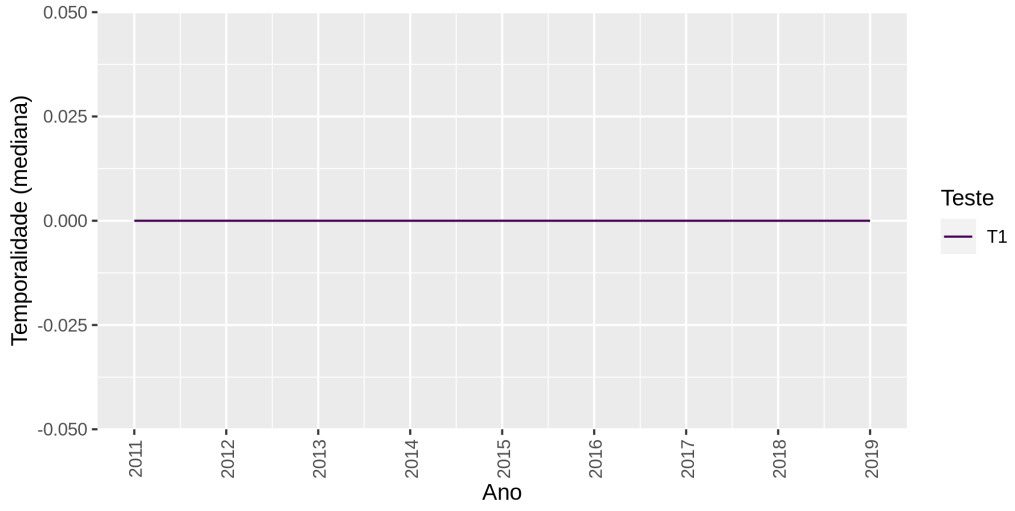

Gráfico 6: distribuição temporal da mediana de temporalidade.

5 Considerações finais

A avaliação realizada é especialmente oportuna, tendo em vista o cenário nacional e o atual empenho em fomentar o debate em torno da qualidade das informações sobre a linha da vida do brasileiro.

Analisado os resultados obtidos pelas métricas de completude, conformidade e acurácia, grande parte das variáveis analisadas apresentaram bons resultados gerais. Menciona-se a ocorrência, mesmo que em pequena escala, do caractere espaço, que pode representar um problema visto que estará de acordo ao tamanho estabelecido pelo dicionário de dados, porém sem representar informação alguma, além da presença de registros ignorados.

Com relação aos resultados de consistência, conclui-se que o ano de 2011 registrou maior quantidade de inconsistências em relação aos demais e o ano de 2018 registrou a menor quantidade de inconsistências. De forma geral, a base avaliada é bastante consistente, visto que foram realizados 5 testes de inconsistência, em dois testes o resultado obtido foi 0 e os demais testes ficaram próximos ao mesmo.

Acerca dos resultados de unicidade, conclui-se que existem registros que apresentam inconsistências entre informações de um mesmo paciente. No geral 0,66% dos identificadores registram pelo menos um dos erros. Sobre a distribuição temporal, conclui-se que houve um bom resultado em relação aos anos, sendo o menor deles 99,66%.

Sobre os resultados de temporalidade, foi possível observar variações nos testes realizados. Os valores máximos e mínimos encontrados demonstram demora para realização do processo executado, ou até mesmo o registro incorreto da informação.

Finalmente, a média ponderada dos resultados de completude é 89,22%, de conformidade é 100,00%, de acurácia é 100,00%, de consistência é 99,97% e de unicidade é 99,57%. Realizando o produto destes resultados, obtêm-se 88,81%, caracterizando a **base de dados como ótima**.

Referências

1 MERINO, J. et al. A data quality in use model for big data. *Future Generation Computer Systems*, v. 63, p. 123–130, 2016. ISSN 0167-739X. Modeling and Management for Big Data Analytics and Visualization. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0167739X15003817>.

2 DAMA, U. The six primary dimensions for data quality assessment-defining data quality dimensions. *Bristol: np URL: https://www.whitepapers.em360tech.com/wp-content/files_mf/1407250286DAMAUKDQDimensionsWhitePaperR37.pdf*, v. 3, p. 2017, 2013.

3 SAúDE, M. da. *HÓRUS*. 2020. Online; acessado em 22 de junho de 2020. Disponível em: <https://www.saude.gov.br/assistencia-farmaceutica/sistema-horus>.

4 SAúDE, B. M. da Saúde. Secretaria de Vigilância em Saúde. Departamento de Análise de Situação em. *Saúde Brasil 2019: uma análise da situação de saúde com enfoque nas doenças imunopreveníveis e na imunização*. [S.l.]: Ministério da Saúde, 2019.

Apêndice A Descrição das variáveis

Variável	Descrição	Tamanho
co_concentracao_medic		6
co_formula_farmaceutica		6
co_municipio_ibge	Município do paciente e da unidade de saúde	6
co_principio_ativo_medicam		6
co_produto		6
co_seq_paciente	Sequencial de identificação do paciente	12
co_tipo_produto		6
co_unidade_consumo		9
co_unidade_medida_item	Código da unidade de medida do medicamento	3
co_unidade_medida_produto		3
co_volume_medicamento		6
ds_concentracao_medicamento	Descrição da concentração do medicamento	200
ds_forma_farmaceutica	Descrição da forma farmacêutica (suspensão oral, comprimido, cápsula, pó para suspensão, xarope etc)	200
ds_principio_ativo_medicamento	Descrição do princípio ativo	200
ds_produto	Descrição do produto (exemplo: omeprazol 20 mg cápsula)	250
ds_tipo_produto	Descrição do tipo de medicamento	50
dt_atendimento	Data do atendimento ao paciente	
dt_nascimento	Data de nascimento do paciente	
dt_receita	Data de emissão da receita (para pegar medicamento na farmácia)	
id_paciente	Identificador do paciente	
nu_concentracao_medic		50
nu_dias_dispensar	Dias de uso	3
nu_fator_correcao		6
nu_produto		20
nu_volume_medicamento		50
qt_dose	Quantidade de doses	6
qt_duracao_tratam_dia	Quantidade de dias do tratamento com o medicamento	3
qt_posologia	Posologia	4

Variável	**Descrição**	**Tamanho**
qt_solicitado	Quantidade solicitada	6
sg_tipo_produto		1
st_rename		1
tp_produto		1
tp_sexo	Sexo do paciente	1
vl_item_dispensacao		30

Apêndice B Resultados numéricos

B.1 Resultados gerais

Variável	Completude [%]	Conformidade [%]	Acurácia [%]
co_concentracao_medic	15,37	100,00	100,00
co_formula_farmaceutica	94,94	100,00	100,00
co_municipio_ibge	100,00	100,00	100,00
co_principio_ativo_medicam	94,94	100,00	100,00
co_produto	99,93	100,00	100,00
co_seq_paciente	100,00	100,00	100,00
co_tipo_produto	99,93	100,00	100,00
co_unidade_consumo	98,85	100,00	100,00
co_unidade_medida_item	99,78	100,00	100,00
co_unidade_medida_produto	99,93	100,00	100,00
co_volume_medicamento	19,24	100,00	100,00
ds_concentracao_medicamento	15,37	100,00	100,00
ds_forma_farmaceutica	94,94	100,00	100,00
ds_principio_ativo_medicamento	94,94	100,00	100,00
ds_produto	99,93	100,00	100,00
ds_tipo_produto	99,93	100,00	100,00
dt_atendimento	100,00	100,00	100,00
dt_nascimento	100,00	100,00	100,00
dt_receita	99,78	100,00	100,00
id_paciente	98,20	100,00	100,00
nu_concentracao_medic	93,14	100,00	100,00
nu_dias_dispensar	99,93	100,00	100,00
nu_fator_correcao	99,36	100,00	100,00
nu_produto	99,93	100,00	100,00
nu_volume_medicamento	19,40	100,00	99,93
qt_dose	99,93	100,00	100,00
qt_duracao_tratam_dia	99,93	100,00	100,00
qt_posologia	99,93	100,00	100,00

Variável	Completude [%]	Conformidade [%]	Acurácia [%]
qt_solicitado	99,93	100,00	100,00
sg_tipo_produto	99,93	100,00	100,00
st_rename	99,93	100,00	100,00
tp_produto	99,93	100,00	100,00
tp_sexo	100,00	100,00	100,00
vl_item_dispensacao	96,45	100,00	100,00

B.2 Resultados por ano

Ano	Completude [%]	Conformidade [%]	Acurácia [%]
2011	89,40	100,00	100,00
2012	89,12	100,00	100,00
2013	89,19	100,00	100,00
2014	88,93	100,00	100,00
2015	89,32	100,00	100,00
2016	89,50	100,00	100,00
2017	89,41	100,00	100,00
2018	89,38	100,00	100,00
2019	88,85	100,00	100,00

Apêndice C Valores atípicos

Variável	Valores atípicos
nu_dias_dispensar	61, 62, 63, 64, 65, 66, 67, 68, 69, 70, ...[3], 754, 770, 800, 820, 830, 840, 870, 900, 990, 999
nu_fator_correcao	2, 5, 8, 10, 14, 15, 20, 21, 22, 25, ...[3], 473, 480, 500, 560, 600, 750, 800, 900, 1000, 1200
qt_dose	0.1, 0.11, 0.12, 0.13, 0.14, 0.15, 0.16, 0.17, 0.18, 0.19, ...[3], 807, 840, 850, 864, 900, 901, 930, 960, 990, 999
qt_duracao_tratam_dia	61, 62, 63, 64, 65, 66, 67, 68, 69, 70, ...[3], 840, 870, 890, 900, 901, 902, 910, 920, 990, 999
qt_posologia	4, 4.3, 4.5, 5, 6, 7, 7.5, 8, 9, 10, ...[3], 300, 310, 330, 350, 355, 360, 401, 430, 464, 500
qt_solicitado	136, 137, 138, 139, 140, 141, 142, 143, 144, 145, ...[3], 3e+05, 324000, 330750, 345000, 359640, 390000, 432000, 518400, 546750, 864000
vl_item_dispensacao	-881888.657, -440933.419, -249754.04, -124866.111, -33757.925, -202.428, -53.532, -51.031, -41.827, -36.303, ...[3], 15833333.377, 27069934.764, 31666666.515, 49241492.189, 92840721.21, 185681432.419, 371362804.837, 742725579.674, 1022245955.358, 2044491895.715

[3] Os valores foram comprimidos devido a alta quantidade.

Apêndice D Testes de inconsistências

D.1 Testes realizados

- **Teste 1:** A data de emissão do paciente não pode ser maior que a data de atendimento do paciente;
- **Teste 2:** A data de nascimento do paciente não pode ser maior que a data de emissão do paciente;
- **Teste 3:** A data de nascimento do paciente não pode ser maior que a data de atendimento do paciente;
- **Teste 4:** Se a descrição do tipo de produto for *MEDICAMENTO*, a sigla do tipo de produto precisa ser *M*;
- **Teste 5:** Se a descrição do tipo de produto for *PRODUTO PARA SAÚDE*, a sigla do tipo de produto precisa ser *I*.

D.2 Resultados obtidos

Ano	**Teste 1**	**Teste 2**	**Teste 3**	**Teste 4**	**Teste 5**
2011	2402	140	104	0	0
2012	5106	202	74	0	0
2013	1587	201	6	0	0
2014	90	47	26	0	0
2015	93	13	6	0	0
2016	41	62	4	0	0
2017	84	28	0	0	0
2018	106	41	5	0	0
2019	68	70	4	0	0