PROADI - Hospital Israelita Albert Einstein

Versão gerada automaticamente

PROADI - Hospital Israelita Albert Einstein

Qualidade de dados

SIA-RAAS-PSI

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 (Merino et al. 2016). 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. 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 (Merino et al. 2016).

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. 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. Nesse contexto, as seguintes dimensões de qualidade de dados são analisadas: Completude, Conformidade, Acurácia, Consistência, Unicidade, Temporalidade

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, por exemplo, o nascimento de um recém-nascido e inclusão desse registro no sistema. O principal interesse é verificar se o dado é disponibilizado prontamente.

2 Métodos

A análise apresentada 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. Nesse contexto, são definidos parâmetros a serem passados para as funções relativas às metricas citadas e implementação de *queries* para os testes de consistência, sintetizados em um único *script* relativo à base analisada.

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 em um 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, é inferido das descrições das variáveis contidas nos relatórios de integração e é apresentado neste relatório. No que tange os testes de unicidade, procurou-se analisar apenas as informações individuais dos pacientes.

Mudanças no domínio e tamanho de caracteres das variáveis são detectadas, relatadas e consideradas no cálculo de medidas de qualidade dos dados.

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, sendo que 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 (Brasil. Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Análise de Situação em Saúde 2019). Em decorrência do método cascata utilizado, é realizado o produto dos resultados obtidos, caracterizando a qualidade da base de dados como um todo, que também pode ser classificada considerando as classes definidas em *Saúde Brasil* (Brasil. Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Análise de Situação em Saúde 2019).

3 Disponibilidade dos dados

Esta análise têm o objetivo de dissertar acerca da disponibilidade dos dados em todo o período representado pela base de dados e em todas as Unidades Federativas.

Após realização de testes, averiguou-se que para os seguintes anos, meses ou Unidades Federativas, representados na tabela a seguir, os respectivos registros não encontram-se na base de dados.

Tabela 1: período e/ou Unidade Federativa contendo registros faltantes.

Ano	Mês
2019	12
2019	10
2019	7
2019	9
2019	8
2019	11

4 Variáveis existentes e mudanças ocorridas

Esta análise tem o objetivo identificar as variáveis existentes na base de dado se relatar as mudanças ocorridas ao longo do tempo.

Após realização de testes não identificou-se qualquer alteração significante nas variáveis.

5 Resultados

Descrições das variáveis são apresentadas no Dicionário adotado. Resultados dos testes de completude, conformidade e acurácia são exibidos nos Resultados numéricos, onde estão organizados em duas tabelas: resultado geral e resultado agregado por ano.

5.1 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, e `NULL`, que representa objetos nulos.

No geral, os resultados de completude das variáveis estão distribuídas pelas categorias definidas em Métodos segundo o gráfico a seguir. O resultado percentual por variável está descrito nos Resultados numéricos. O cômputo da média ponderada dos resultados obtidos é de **75.93%**, ou seja, a **completude é ótimo**.

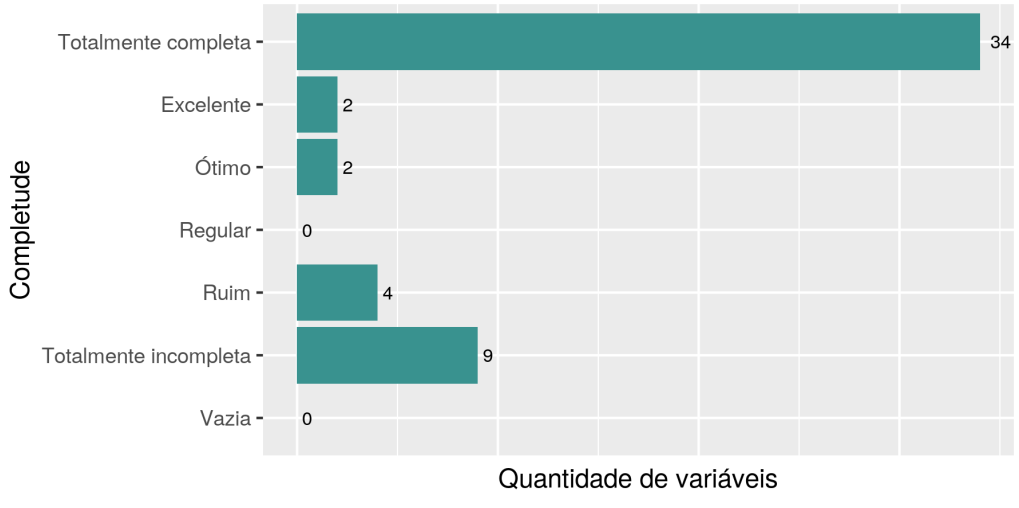

Gráfico 1: distribuição dos resultados de completude.

5.2 Conformidade

Verificou-se se os dados apresentam os padrões descritos no dicionário de dados adotado como referência a respeito da quantidade de caracteres e valores válidos.

Ressalta-se que durante a construção do dicionário de dados não foi possível obter os microdados ou informação equivalente contendo, quando existente, os valores válidos de domínio para algumas variáveis. Nesse sentido, a tabela a seguir apresenta ao máximo dez registros mais frequentes para essas variáveis, separados por vírgulas.

Tabela 2: registros mais frequentes por variável em que não foi possível obter a descrição.

Variável	Domínio

No geral, os resultados de conformidade das variáveis estão distribuídas pelas categorias definidas em Métodos segundo o gráfico a seguir. O resultado percentual por variável está descrito nos Resultados numéricos. O cômputo da média ponderada dos resultados obtidos é de **97.99%**, ou seja, a **conformidade é excelente**.

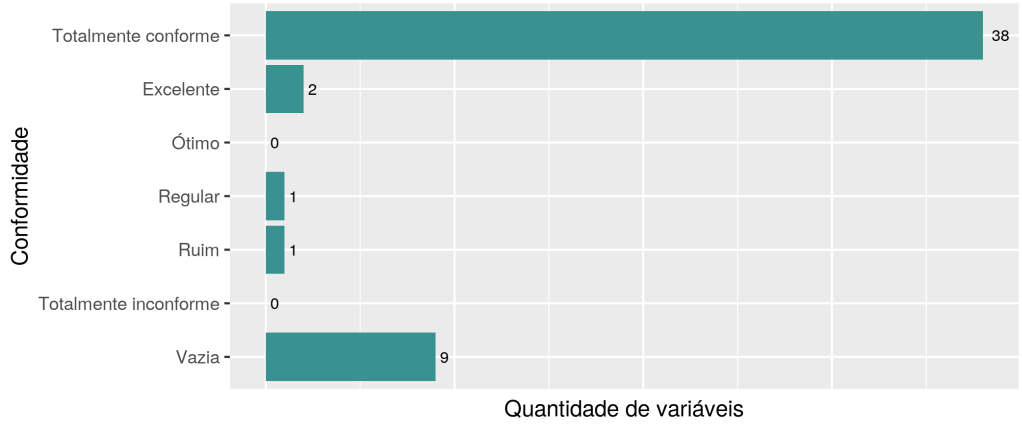

Gráfico 2: distribuição dos resultados de conformidade.

5.3 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[1]. Após esta análise de datas e municípios é realizada investigação acerca do preenchimento

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

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).

No que tange propriamente o preenchimento dos registros, buscou-se identificar valores representando a ausência de informações ou que foram ignorados, além de sequências finitas do caractere espaço (*whitespace*) ou sequências finitas do numeral zero para variáveis não numéricas. Este fato pode representar um problema, visto que estará de acordo ao tamanho estabelecido pelo dicionário de dados, porém não estará acurado, não representando informação alguma. Para realizar a identificação no primeiro caso, utilizou-se o método TF-IDF.

O método TF-IDF[2], em conjunto aos métodos *N-Grams* e multiplicação de matriz esparsa, é uma medida estatística que tem o intuito de indicar a similaridade de uma palavra em relação a outra. TF-IDF é um método para gerar recursos do texto multiplicando a frequência de um termo em um documento (*Term Frequency*, ou TF) pela importância (*Inverse Document Frequency*, ou IDF) do mesmo termo em um corpus inteiro. Este método é muito útil na classificação e no agrupamento de textos e é usado para transformar documentos em vetores numéricos, que podem ser facilmente comparados. Embora os termos no TF-IDF sejam geralmente palavras, essa condição não é necessária. Como a maioria dos registros possuem de uma a três palavras, utilizou-se *N-Grams*: sequências de N caracteres contíguos. Para avaliação, calculou-se a proximidade dos vetores resultantes do método TF-IDF, através da semelhança cosseno, que pode ser vista como um produto escalar normalizado.

Após aplicação do método, não identificou-se qualquer registro nessa situação.

No geral, os resultados de acurácia das variáveis estão distribuídas pelas categorias definidas em Métodos segundo o gráfico a seguir. O resultado percentual por variável está descrito nos Resultados numéricos. O cômputo da média ponderada dos resultados obtidos é de **99.23%**, ou seja, a **acurácia é excelente**.

[2] https://bergvca.github.io/2017/10/14/super-fast-string-matching.html

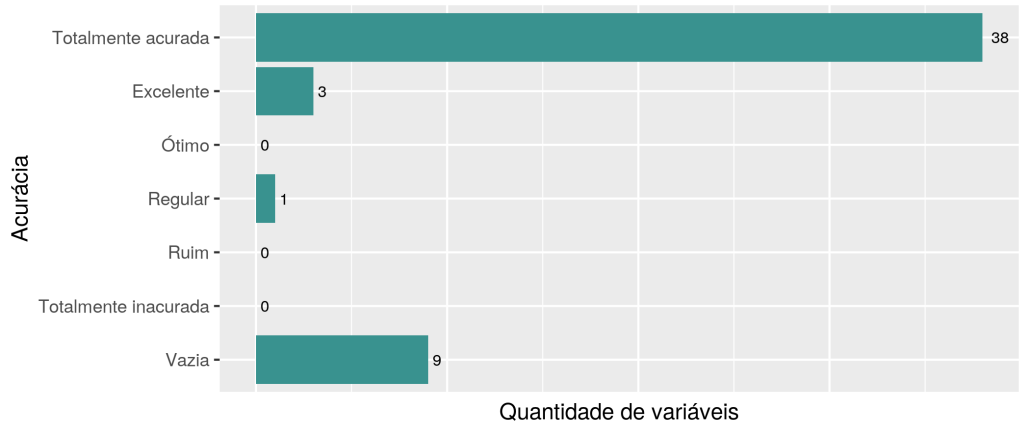

Gráfico 3: distribuição dos resultados de acurácia.

5.4 Consistência

Os resultados apresentados 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. Uma descrição mais detalhada de cada teste está presente em Testes de inconsistência.

Tabela 3: resultados de consistência.

Teste	Descrição	Falhas [partes por mil]
T1	dt_realizacao > dt_fim	0.000
T2	dt_realizacao < dt_inicio	0.015
T3	dt_fim < dt_inicio	0.000
T4	dt_nascimento > dt_inicio	0.000
T5	dt_nascimento > dt_fim	0.000
T6	dt_nascimento > dt_realizacao	0.000

A distribuição temporal dos resultados dos testes de consistência é apresentada no gráfico a seguir

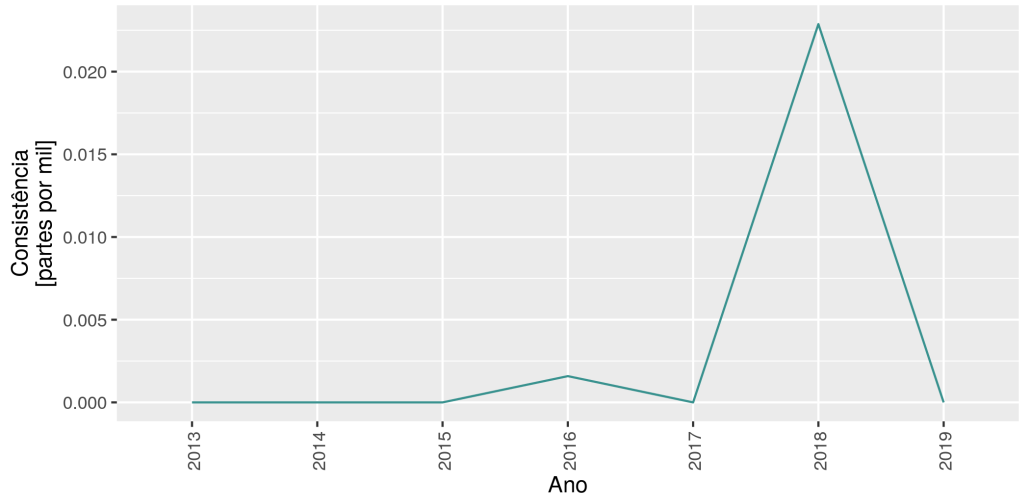

Gráfico 4: distribuição temporal da consistência.

O resultado de cada teste por ano está descrito em Testes de inconsistência. O cômputo da média ponderada dos resultados obtidos é de **100.00%**, ou seja, a **consistência é excelente**.

5.5 Unicidade

Nesta dimensão é calculado o grau de duplicidade dos dados, buscando diferenças por meio dos identificadores dos pacientes.

São excluídos deste teste os identificadores de pacientes presentes apenas uma vez na base de dados.

Tabela 4: resultados de unicidade por variável relacionada à identificação do paciente.

Teste	Variável	Unicidade [%]
T1	*dt_nascimento*	94.02
T2	*ds_sexo*	97.61

A distribuição temporal dos resultados dos testes de unicidade é apresentada no gráfico a seguir.

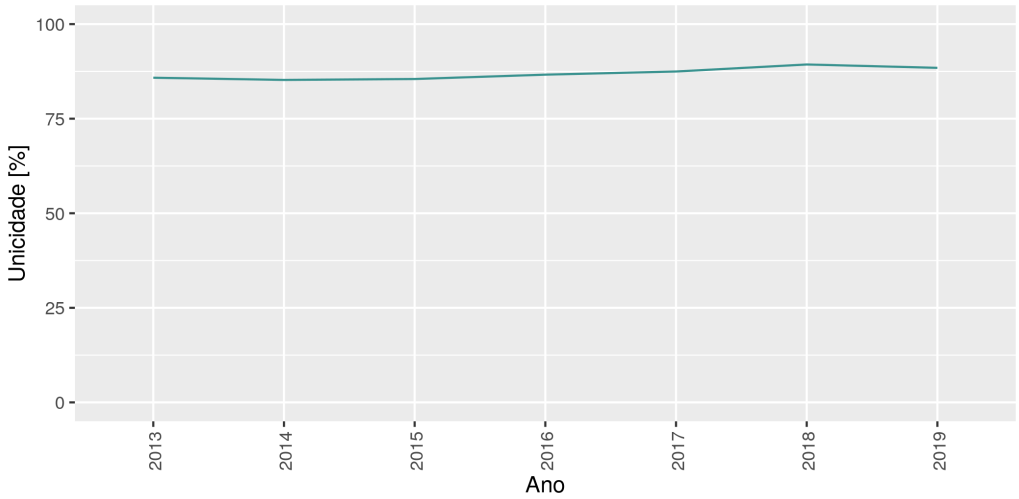

Gráfico 5: distribuição temporal da unicidade.

O cômputo da média ponderada dos resultados obtidos é de **92.04%**, ou seja, a **unicidade é excelente**.

5.6 Temporalidade

Para mensurar esta dimensão é calculada a quantidade de dias entre duas variáveis representando datas que estejam conformes, acuradas e consistentes.

Tabela 5: resultados de temporalidade.

Teste	Variável Inicial	Variável final	Mediana	Min.	Max.
T1	*dt_inicio*	*dt_fim*	810	0	8100
T2	*dt_inicio*	*dt_competencia*	1500	31	29000
T3	*dt_competencia*	*dt_fim*			

A distribuição temporal das medianas obtidas pelos testes de temporalidade é apresentada no gráfico a seguir.

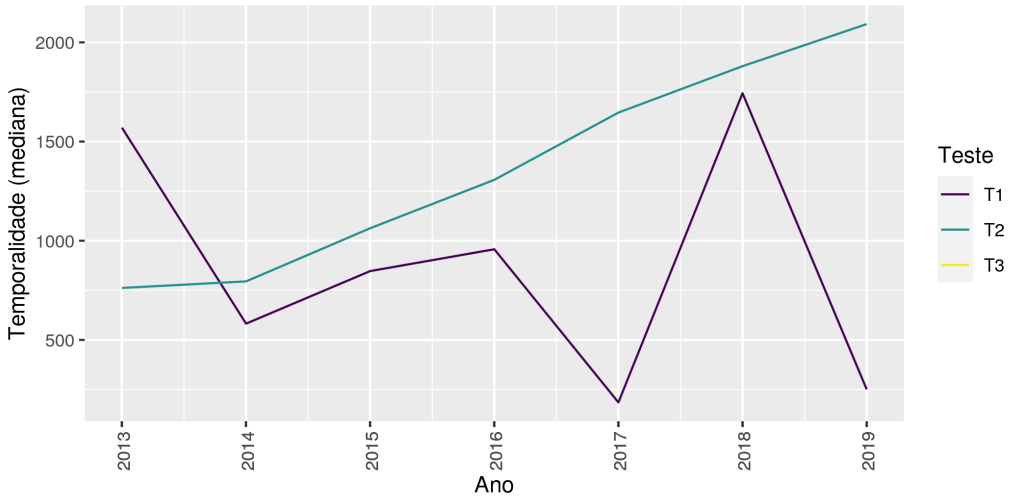

Gráfico 6: distribuição temporal da temporalidade.

6 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 acerca de estabelecimentos de saúde do país.

Assim, a média ponderada dos resultados de Completude é 75.93%, de Conformidade é 97.99%, de Acurácia é 99.23%, de Consistência é 100.00%, de Unicidade é 92.04%. Realizando o produto destes resultados, obtêm-se **67.96%**, caracterizando a **base de dados como regular**.

7 Referências

Brasil. Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Análise de Situação em Saúde. 2019. *Saúde Brasil 2019: Uma Análise Da Situação de Saúde Com Enfoque Nas Doenças Imunopreveníveis E Na Imunização*. Ministério da Saúde.

Merino, Jorge, Ismael Caballero, Bibiano Rivas, Manuel Serrano, and Mario Piattini. 2016. "A Data Quality in Use Model for Big Data." *Future Generation Computer Systems* 63: 123–30. https://doi.org/https://doi.org/10.1016/j.future.2015.11.024.

Dicionário adotado

Variável	Descrição	Tamanho
co_autorizacao		[1, 13]
co_cbo		[6, 6]
co_celular		[11, 11]
co_cep		[8, 8]
co_cid_causas_assoc		[1, 4]
co_cid_principal		[1, 4]
co_cid_procedimento		[1, 6]
co_classificacao		[1, 3]
co_cnes_esf		[7, 7]
co_cnes_executante		[7, 7]
co_cns_medico		[32, 32]
co_cns_paciente		[15, 15]
co_cobertura_esf		[1, 1]
co_etnia		[1, 4]
co_gestor		[6, 6]
co_inst_registro		[1, 3]
co_municipio		[6, 6]
co_nacionalidade		[1, 3]
co_procedimento		[1, 10]
co_raca		[1, 2]
co_remessa		[1, 22]
co_servico		[1, 3]
co_telefone		[11, 11]
ds_complemento		[1, 10]
ds_destino		[1, 22]
ds_local_realizacao		[1, 10]
ds_logradouro		[1, 30]
ds_mae		[1, 30]
ds_origem		[1, 6]
ds_paciente		[1, 30]
ds_prontuario		[1, 10]
ds_responsavel		[1, 30]
ds_sexo		[1, 1]
dt_competencia		[6, 6]
dt_fim		[8, 8]
dt_inicio		[8, 8]
dt_nascimento		[8, 8]
dt_realizacao		[8, 8]
flg_usuario_droga		[1, 1]
id_paciente		[1, 30]
no_cid_causas_associadas		[1, 100]
no_cid_principal		[1, 100]
no_fantasia		[1, 60]
no_fantasia_esf		[1, 60]
no_municipio		[1, 60]
no_procedimento		[1, 250]
no_razao_social		[1, 60]

(continued)

Variável	Descrição	Tamanho
no_razao_social_esf		[1, 60]
qt_pprocedimento		[1, 12]
sg_uf		[2, 2]
tp_droga		[1, 21]

Resultados numéricos

Resultados gerais

Variável	Completude [%]	Conformidade [%]	Acurácia [%]
co_autorizacao	87.05	100.00	100.00
co_cbo	100.00	100.00	100.00
co_celular	0.00	0.00	0.00
co_cep	0.00	0.00	0.00
co_cid_causas_assoc	100.00	100.00	100.00
co_cid_principal	100.00	100.00	100.00
co_cid_procedimento	100.00	100.00	100.00
co_classificacao	100.00	100.00	100.00
co_cnes_esf	100.00	69.52	68.57
co_cnes_executante	100.00	100.00	100.00
co_cns_medico	100.00	100.00	100.00
co_cns_paciente	0.00	0.00	0.00
co_cobertura_esf	100.00	100.00	100.00
co_etnia	100.00	100.00	100.00
co_gestor	100.00	100.00	100.00
co_inst_registro	100.00	100.00	100.00
co_municipio	100.00	100.00	100.00
co_nacionalidade	100.00	100.00	100.00
co_procedimento	100.00	100.00	100.00
co_raca	100.00	100.00	100.00
co_remessa	100.00	100.00	100.00
co_servico	100.00	100.00	100.00
co_telefone	0.00	0.00	0.00
ds_complemento	0.00	0.00	0.00
ds_destino	100.00	99.98	100.00
ds_local_realizacao	100.00	100.00	100.00
ds_logradouro	0.00	0.00	0.00
ds_mae	0.00	0.00	0.00
ds_origem	5.66	100.00	100.00
ds_paciente	0.00	0.00	0.00
ds_prontuario	92.74	100.00	99.22
ds_responsavel	0.00	0.00	0.00
ds_sexo	100.00	100.00	100.00
dt_competencia	100.00	100.00	100.00
dt_fim	100.00	45.94	100.00
dt_inicio	100.00	100.00	100.00
dt_nascimento	100.00	100.00	100.00
dt_realizacao	100.00	100.00	100.00
flg_usuario_droga	87.05	100.00	100.00
id_paciente	93.37	100.00	100.00
no_cid_causas_associadas	11.45	100.00	100.00
no_cid_principal	100.00	100.00	100.00
no_fantasia	100.00	100.00	100.00
no_fantasia_esf	47.65	100.00	100.00
no_municipio	100.00	100.00	100.00

(continued)

Variável	Completude [%]	Conformidade [%]	Acurácia [%]
no_procedimento	100.00	100.00	100.00
no_razao_social	100.00	100.00	100.00
no_razao_social_esf	47.65	100.00	100.00
qt_pprocedimento	100.00	100.00	99.97
sg_uf	100.00	100.00	100.00
tp_droga	100.00	100.00	100.00

Resultados por ano

Ano	Completude [%]	Conformidade [%]	Acurácia [%]
2013	100.00	100.00	98.75
2014	100.00	100.00	98.81
2015	100.00	100.00	98.78
2016	100.00	97.94	99.30
2017	100.00	95.96	99.98
2018	100.00	96.05	99.96
2019	100.00	95.93	99.95

Testes de inconsistência

Testes realizados

- **T1**: A data de realização não pode ser maior que a data final
- **T2**: A data de realização não pode ser menor que a data de inicio
- **T3**: A data final não pode ser menor que a data inicial
- **T4**: A data de nascimento não pode ser maior que a data de início
- **T5**: A data de nascimento não pode ser maior que a data de fim
- **T6**: A data de nascimento não pode ser maior que a data de realização

Resultados obtidos

Ano	T1	T2	T3	T4	T5	T6
2013	0	0	0	0	0	0
2014	0	0	0	0	0	0
2015	0	0	0	0	0	0
2016	0	1	0	0	0	0
2017	0	0	0	0	0	0
2018	0	21	0	0	0	0
2019	0	0	0	0	0	0