

PROADI - Hospital Israelita Albert Einstein

Versão gerada automaticamente

PROADI - Hospital Israelita Albert Einstein

Qualidade de dados

SIA-BPAI

Sumário

1 Qualidade de dados	4
2 Métodos	5
3 Disponibilidade dos dados	6
4 Variáveis existentes e mudanças ocorridas	6
5 Resultados	7
5.1 Completude	7
5.2 Conformidade	8
5.3 Acurácia	8
5.4 Unicidade	10
5.5 Temporalidade	11
6 Considerações finais	12
7 Referências	13
Dicionário adotado	14
Resultados numéricos	15
Resultados gerais	15
Resultados por ano	15
Resultados obtidos	16

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, 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.

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 métricas citadas e implementação de *queries* para os testes de consistência, sintetizados em um único *script* relativo à base analisada.

O manejo 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 submeti-

dos 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 tem 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	11
2019	10

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 significativa 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. Uma descrição mais detalhada dos testes de inconsistência realizados, bem como seus respectivos resultados numéricos estão descritos em Testes de inconsistência.

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 **92.52%**, ou seja, a **completude é excelente**.

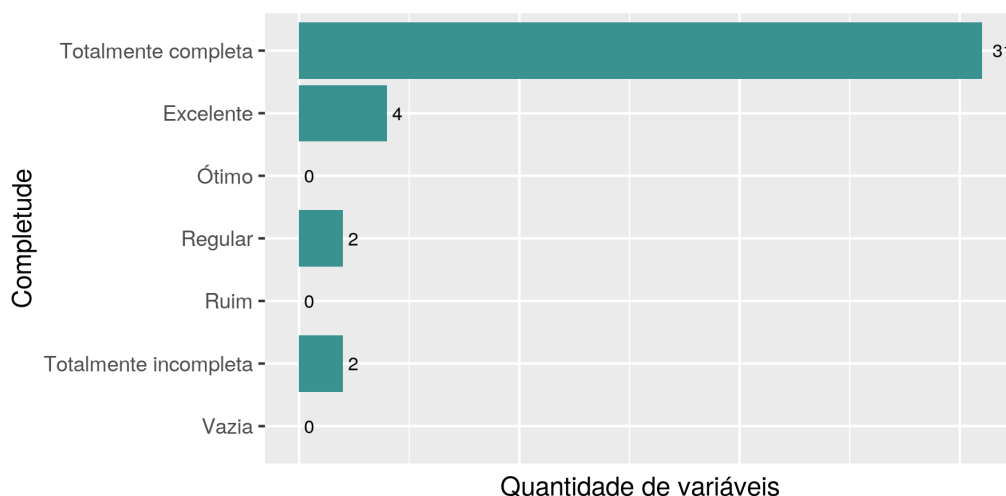


Gráfico 1: distribuição dos resultados de completude.

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.

<i>Variável</i>	<i>Domínio</i>
-----------------	----------------

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 **99.99%**, ou seja, a **conformidade é excelente**.

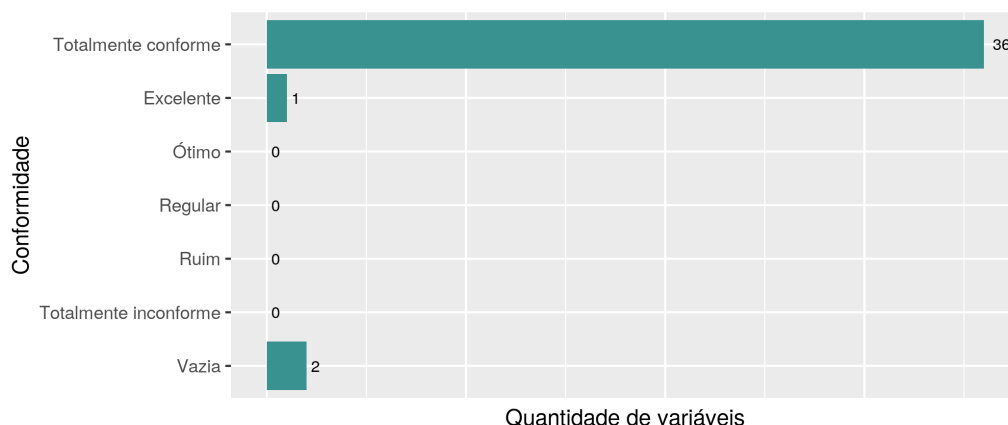


Gráfico 2: distribuição dos resultados de conformidade.

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 IBGE¹. 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-IDF², 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.

Em relação as várias quantitativas, a tabela a seguir expõe os valores atípicos detectados³, isto é, registros numéricos que apresentam grande afastamento em relação aos demais, dentro do universo de uma única variável, os quais implicam, tipicamente, em prejuízos a interpretação dos resultados dos testes estatísticos aplicados.

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

³<https://www.rdocumentation.org/packages/grDevices/versions/3.6.2/topics/boxplot.stats>

PROADI - Hospital Israelita Albert Einstein

Tabela 3: valores atípicos para variáveis numéricas.

Variável	Valores atípicos
<i>idade</i>	107, 108, 109, 110, 111, 112, 113, 114, 115, 116, ..., 120, 121, 122, 123, 125, 126, 127, 128, 129, 130
<i>qt_realizado</i>	4, 5, 6, 7, 8, 9, 10, 11, 12, 13, ..., 1010, 1055, 1111, 1116, 1231, 1301, 1500, 2801, 4610, 8715
<i>valor</i>	111.3, 111.96, 112, 112.66, 112.7, 113.26, 113.68, 113.85, 114.3, 114.36, ..., 26317.5, 27621, 28314, 31484.2, 34167.76, 37389, 38521.98, 39007.21, 106297.95, 1581772.5

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 **99.91%**, ou seja, a **acurácia é excelente**.

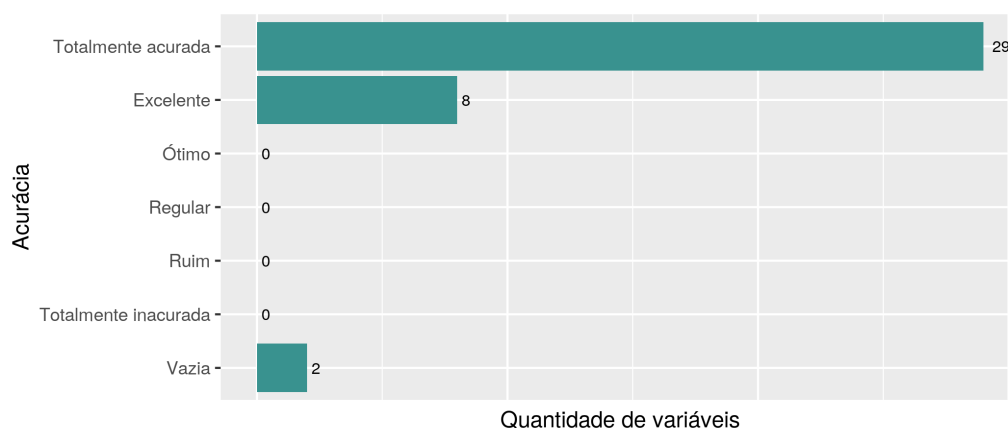


Gráfico 3: distribuição dos resultados de acurácia.

5.4 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 [%]
<i>T1</i>	<i>dt_nascimento</i>	99.77
<i>T2</i>	<i>sexo</i>	99.07

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

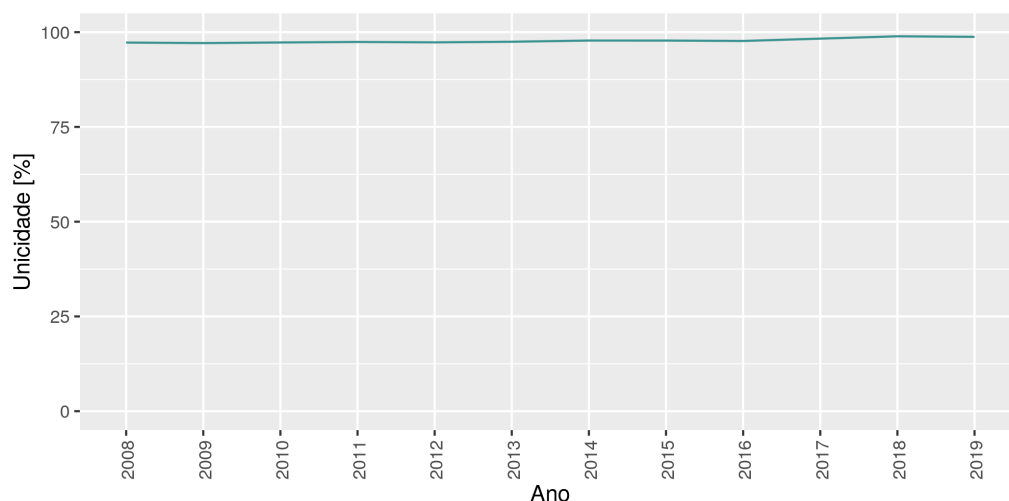


Gráfico 4: distribuição temporal da unicidade.

O cômputo da média ponderada dos resultados obtidos é de **98.85%**, ou seja, a **unicidade é excelente**.

5.5 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.
<i>T1</i>	<i>dt_atendimento</i>	<i>competencia</i>	52	31	4200

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

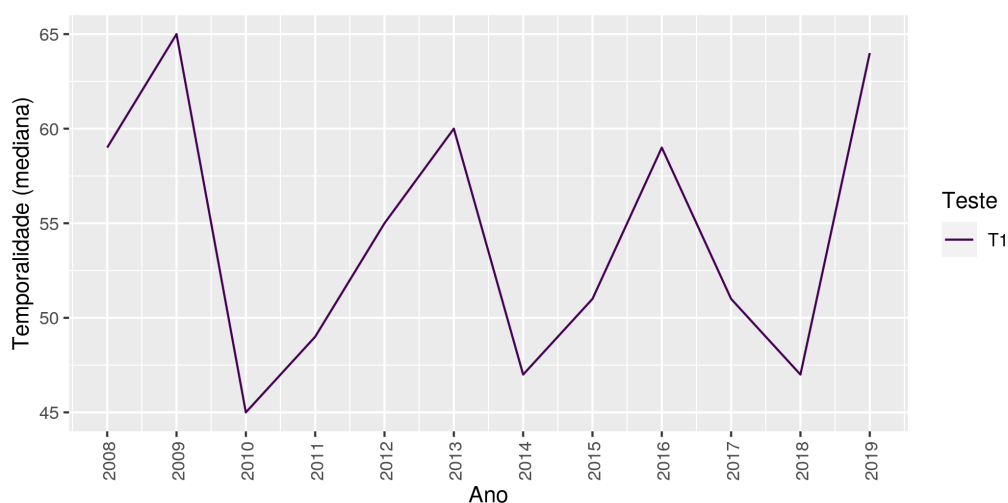


Gráfico 5: distribuição temporal da temporalidade.

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 92.52%, de Conformidade é 99.99%, de Acurácia é 99.91%, de Unicidade é 98.85%. Realizando o produto destes resultados, obtêm-se **91.37%**, caracterizando a **base de dados como excelente**.

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
<i>cbo</i>		[6, 6]
<i>cnes</i>		[7, 7]
<i>cns_medico</i>		[32, 32]
<i>cns_paciente</i>		[15, 15]
<i>co_bpi_mvm</i>		[1, 6]
<i>co_bpi_tpfm</i>		[1, 1]
<i>co_carater_atendimento</i>		[1, 2]
<i>co_cid_principal</i>		[1, 4]
<i>co_municipio</i>		[6, 6]
<i>co_origem</i>		[1, 3]
<i>co_procedimento_realizado</i>		[1, 10]
<i>co_rms</i>		[1, 4]
<i>competencia</i>		[6, 6]
<i>condicao</i>		[1, 2]
<i>desc_cbo</i>		[1, 200]
<i>ds_tp_financiamento</i>		[1, 4]
<i>dt_atendimento</i>		[8, 8]
<i>dt_nascimento</i>		[8, 8]
<i>folha</i>		[1, 3]
<i>gestor</i>		[6, 6]
<i>id_paciente</i>		[1, 30]
<i>idade</i>		[1, 3]
<i>municipio_paciente</i>		[6, 6]
<i>no_carater_atendimento</i>		[1, 255]
<i>no_cid_principal</i>		[1, 100]
<i>no_fantasia</i>		[1, 35]
<i>no_municipio</i>		[1, 60]
<i>no_municipio_paciente</i>		[1, 60]
<i>no_paciente</i>		[1, 30]
<i>no_procedimento_realizado</i>		[1, 250]
<i>no_razao_social</i>		[1, 35]
<i>nu_autorizacao</i>		[1, 13]
<i>qt_realizado</i>		[1, 8]
<i>remessa</i>		[1, 21]
<i>sequencial</i>		[1, 2]
<i>sexo</i>		[1, 1]
<i>uf</i>		[2, 2]
<i>uf_paciente</i>		[2, 2]
<i>valor</i>		[1, 12]

PROADI - Hospital Israelita Albert Einstein

Resultados numéricos

Resultados gerais

Variável	Compleitude [%]	Conformidade [%]	Acurácia [%]
cbo	100.00	100.00	100.00
cnes	100.00	100.00	100.00
cns_medico	100.00	100.00	100.00
cns_paciente	0.00	0.00	0.00
co_bpi_mvm	100.00	100.00	100.00
co_bpi_tpin	100.00	100.00	100.00
co_carater_atendimento	100.00	100.00	100.00
co_cid_principal	100.00	100.00	100.00
co_municipio	100.00	100.00	100.00
co_origem	100.00	100.00	100.00
co_procedimento_realizado	100.00	100.00	100.00
co_rms	100.00	100.00	100.00
competencia	100.00	100.00	100.00
condicao	100.00	100.00	100.00
desc_cbo	100.00	100.00	100.00
ds_tp_financiamento	100.00	100.00	100.00
dt_atendimento	100.00	99.81	100.00
dt_nascimento	100.00	100.00	100.00
folha	100.00	100.00	100.00
gestor	100.00	100.00	96.59
id_paciente	96.25	100.00	100.00
idade	100.00	100.00	100.00
municipio_paciente	100.00	100.00	100.00
no_carater_atendimento	96.52	100.00	100.00
no_cid_principal	52.81	100.00	100.00
no_fantasia	100.00	100.00	100.00
no_municipio	100.00	100.00	100.00
no_municipio_paciente	100.00	100.00	100.00
no_paciente	0.00	0.00	0.00
no_procedimento_realizado	100.00	100.00	100.00
no_razao_social	100.00	100.00	100.00
nu_autorizacao	62.87	100.00	99.99
qt_realizado	100.00	100.00	100.00
remessa	100.00	100.00	100.00
sequencial	100.00	100.00	100.00
sexo	100.00	100.00	100.00
uf	100.00	100.00	100.00
uf_paciente	100.00	100.00	100.00
valor	99.99	100.00	100.00

Resultados por ano

PROADI - Hospital Israelita Albert Einstein

Ano	Compleitude [%]	Conformidade [%]	Acurácia [%]
2008	100.00	100.00	99.86
2009	100.00	100.00	99.87
2010	100.00	100.00	99.92
2011	100.00	100.00	99.91
2012	100.00	100.00	99.89
2013	100.00	100.00	99.91
2014	100.00	100.00	99.92
2015	100.00	100.00	99.88
2016	100.00	99.99	99.88
2017	100.00	99.99	99.89
2018	100.00	99.99	99.92
2019	100.00	99.99	99.93

Resultados obtidos

variavel	total	nao.nulo	conformes	acurados
<i>CBO</i>	16435411	16435411	16435411	16435411
<i>CNES</i>	16435411	16435411	16435411	16435411
<i>CNS_MEDICO</i>	16435411	16435411	16435411	16435411
<i>CNS_PACIENTE</i>	16435411	0	0	0
<i>CO_BPI_MVM</i>	16435411	16435411	16435411	16435411
<i>CO_BPI_TPFIN</i>	16435411	16435411	16435411	16435411
<i>CO_CARATER_ATENDIMENTO</i>	16435411	16435411	16435411	16435371
<i>CO_CID_PRINCIPAL</i>	16435411	16435411	16435411	16435406
<i>CO_MUNICIPIO</i>	16435411	16435411	16435411	16435411
<i>CO_ORIGEM</i>	16435411	16435411	16435411	16435411
<i>CO_PROCEDIMENTO_REALIZADO</i>	16435411	16435411	16435411	16435411
<i>CO_RMS</i>	16435411	16435411	16435411	16435411
<i>COMPETENCIA</i>	16435411	16435411	16435411	16435411
<i>CONDICAO</i>	16435411	16435411	16435411	16435411
<i>DESC_CBO</i>	16435411	16435114	16435114	16435114
<i>DS_TP_FINANCIAMENTO</i>	16435411	16435411	16435411	16435411
<i>DT_ATENDIMENTO</i>	16435411	16435411	16404230	16404230
<i>DT_NASCIMENTO</i>	16435411	16435411	16435411	16434600
<i>FOLHA</i>	16435411	16435411	16435411	16435382
<i>GESTOR</i>	16435411	16435411	16435411	15874631
<i>ID_PACIENTE</i>	16435411	15819540	15819540	15819540
<i>IDADE</i>	16435411	16435411	16435411	16435411
<i>MUNICIPIO_PACIENTE</i>	16435411	16435411	16435411	16435158
<i>NO_CARATER_ATENDIMENTO</i>	16435411	15862804	15862804	15862804
<i>NO_CID_PRINCIPAL</i>	16435411	8679890	8679890	8679890
<i>NO_FANTASIA</i>	16435411	16435411	16435411	16435411
<i>NO_MUNICIPIO</i>	16435411	16435411	16435411	16435411
<i>NO_MUNICIPIO_PACIENTE</i>	16435411	16435411	16435411	16435411
<i>NO_PACIENTE</i>	16435411	0	0	0
<i>NO_PROCEDIMENTO_REALIZADO</i>	16435411	16435411	16435411	16435411
<i>NO_RAZAO_SOCIAL</i>	16435411	16435411	16435411	16435411
<i>NU_AUTORIZACAO</i>	16435411	10332519	10332519	10331392

PROADI - Hospital Israelita Albert Einstein

(continued)

variavel	total	nao.nulo	conformes	acurados
<i>QT_REALIZADO</i>	16435411	16435411	16435411	16435411
<i>REMESSA</i>	16435411	16435411	16435411	16435411
<i>SEQUENCIAL</i>	16435411	16435411	16435411	16435154
<i>SEXO</i>	16435411	16435411	16435411	16435411
<i>UF</i>	16435411	16435411	16435411	16435411
<i>UF_PACIENTE</i>	16435411	16435411	16435411	16435411
<i>VALOR</i>	16435411	16434510	16434510	16434510