PROADI - Hospital Israelita Albert Einstein

Versão preliminar

PROADI - Hospital Israelita Albert Einstein

Qualidade de dados

SIH/SUS-RD

Histórico de revisões

Data	Versão	Descrição	Autor	Responsável
12/06/2020	1.0	Versão preliminar	Leandro Furlam, Elias Ribeiro	Alexandre Rodrigues
07/07/2020	1.0	Inclusão de análise de disponibilidade e mudanças de domínio	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 e temporalidade (2). A dimensão unicidade, que objetiva mensurar o grau de duplicidade nos dados, foi excluída deste estudo, uma vez que dados de identificação dos pacientes são removidos da planilha de informações.

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.

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 expõe a disponibilidade dos dados, a Seção 5 analisa as mudanças que as variáveis sofreram em relação ao tempo, a Seção 6 apresenta os respectivos resultados e a Seção 7 estende a análise com considerações finais.

2 Base de dados

O **Sistema de Informações Hospitalares do SUS (SIH/SUS)**, popularmente conhecido como Sistema AIH, foi criado em 1991 como um instrumento para indução e avaliação das políticas relacionadas à organização e ao financiamento da assistência médico-hospitalar no sistema público de saúde. O SIH foi concebido como um artifício para operar o pagamento das internações e para instrumentalizar ações de controle e auditoria, bem como por utilizado por pesquisadores e gestores. Este sistema abrange apenas a rede pública, e foi desenvolvido e implementado com o objetivo de racionalizar despesas (3).

Operacionaliza através da Autorização de Internação Hospitalar (AIH) (4), e possui como objetivo registrar todos os atendimentos provenientes de internações hospitalares que foram financiadas pelo SUS, e a partir deste processamento, gerar relatórios para que os gestores possam fazer os pagamentos dos estabelecimentos de saúde. Além disso, o nível Federal recebe mensalmente uma base de dados de todas as internações autorizadas (aprovadas ou não para pagamento) para que possam ser repassados às Secretarias de Saúde os valores de Produção de Média e Alta complexidade, além dos valores de CNRAC, FAEC e de Hospitais Universitários, em suas variadas formas de contrato de gestão (3).

Seus alvos são os atendimentos provenientes de internações hospitalares que foram financiadas pelo SUS (4), e dessa forma, o SIH possibilita a avaliação do desempenho e condições sanitárias, através das taxas de óbito e de infecção hos-

pitalar informadas no sistema; fornece informações para a programação do orçamento dos estabelecimentos; gera históricos, permitindo ao Gestor diminuir o volume do banco de produção; otimiza o processamento e a funcionalidade para gerar relatórios a partir do histórico (3).

Os dados são disponibilizados em quatro bases:

- RD – AIH Reduzida;
- RD – AIH Rejeitada;
- RD – AIH Rejeitada com código de erro;
- SP – Serviços Profissionais.

RD – AIH Reduzida contém as AIH aprovadas e também os valores efetivamente pagos por mês de competência. Ela inclui os procedimentos processados e validados pelo Ministério da Saúde entre os apresentados por todos os estabelecimentos prestadores de serviços para o SUS.

RD – AIH Rejeitada contém as AIHs que foram automaticamente bloqueadas pelo SIHSUS no processo de crítica do sistema e que terão que passar pela análise dos gestores/autorizadores que podem confirmar ou não a autorização para a internação.

RD – AIH Rejeitada com código de erro contém as AIHs que foram rejeitadas pelo SIHSUS e que não foram caracterizadas como bloqueio, pois não permitem a liberação pelo gestor. Tais rejeições se devem, em grande parte, a erros de registros na AIH, falta de compatibilidade entre os procedimentos realizados, a informações não compatíveis com o CNES da Instituição, relacionadas aos profissionais, ao credenciamento ou a estrutura física. Neste caso, cabe ao serviço corrigir as informações equivocadas, quando possível, e reapresentar a AIH.

SP – Serviços Profissionais contém as informações dos serviços profissionais (procedimentos hospitalares) realizados no decorrer da internação hospitalar.

2.1 Informações gerais

Base de dados	SIH/SUS-RD
Fonte	<ftp://ftp.datasus.gov.br/dissemin/publicos/SIHSUS/>
Data de obtenção dos dados	12 de junho de 2020
Período	jan/2008 a mar/2020
Região geográfica	Todas as 27 Unidades Federativas
Volume	9,8 GB
Número máximo de variáveis	113
Número de registros	142.579.454

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.

Para analisar a disponibilidade dos dados e as mudanças ocorridas nas variáveis e nos respectivos domínios, informações de diversas fontes, listadas a seguir, foram avaliadas:

- Microdados e as informações obtidas do DATASUS e do Ministério da Saúde, através do Informe Técnico referente ao processamento 2016-03 do Sistema de Informações Hospitalares;
- Arquivos auxiliares e de tabulação, aplicados principalmente no levantamento de domínio e definição de valores ignorados e sem informação;
- Relatório técnico temático produzido pelo Ipea (5), aplicado na análise das mudanças ocorridas nas variáveis ao longo dos anos.

Os resultados apresentados neste relatório consideram a inclusão/retirada de

variáveis ao longo do tempo. Mudanças no domínio das variáveis também 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 e o mesmo se aplica aos registros submetidos aos testes de temporalidade. 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 (6). Em decorrência do método cascata utilizado, é realizado o produto dos resultados obtidos, na Seção 7, caracterizando a qualidade da base de dados como um todo, que também pode ser classificada considerando as classes definidas em *Saúde Brasil* (6).

4 Disponibilidade dos dados

Esta seção 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. Nesse sentido, no Gráfico 1 foram destacados em negrito os anos em que a informação disponibilizada está completa, isto é, os anos que contém todos os meses e Unidades Federativas representadas pela base. Ressalta-se que em 2020 foram disponibilizados até o momento da análise apenas os meses de janeiro, fevereiro e março.

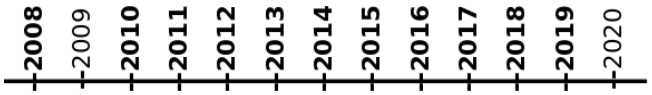

Gráfico 1: disponibilidade dos dados por ano.

Em ambos os anos incompletos, 2009 e 2020, a **ausência ocorre no estado do Acre**. No primeiro caso, não está disponível a informação referente ao mês de setembro; e no segundo caso, a informação referente ao mês de março, o último mês representado pela tabela de dados.

5 Variáveis existentes e mudanças ocorridas

Esta seção tem por objetivo identificar as variáveis existentes na base de dados e relatar as mudanças ocorridas ao longo do tempo. Nesse sentido, o Gráfico 2 apresenta um resumo do quantitativo de variáveis no banco de dados ao longo dos anos analisados, quando podemos verificar que o número de variáveis passou de 86 em 2008 para 113 em 2020. Isso reflete o grande desenvolvimento e avanço de um processo de atualização das informações de saúde para atender à necessidade de informações estatísticas do sistema público de saúde nacional.

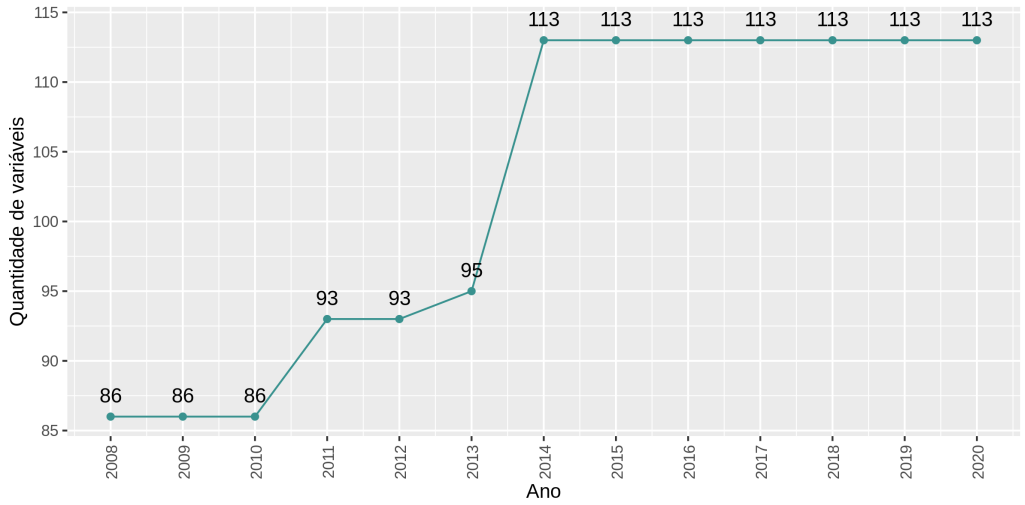

Gráfico 2: evolução do número de variáveis na base de dados.

De fato, há quatro momentos em que houve acréscimo de variáveis: 2011, 2013 e 2014, com todas as alterações ocorrendo no mês de janeiro. As 86 variáveis que já estavam na base de dados foram aumentadas em 7 variáveis no ano de 2011, aumentadas em 2 variáveis no ano de 2013 e novamente aumentadas em 18 variáveis em 2014. A saber, as variáveis existentes por período estão descritas na Tabela 1, separadas por vírgula. Em relação à modificação no domínio de variáveis (variáveis em destaque na Tabela 1), foi elaborada uma descrição detalhada do domínio por ano referente ao tamanho máximo de caracteres permitido para os registros, conforme é descrito na Tabela 2.

Mesmo que presentes na base de dados, identificou-se que algumas variáveis não possuem qualquer registro em um determinado ano, isto é, estão totalmente vazias em um período específico. Esse fato torna-se um problema quando deseja-se realizar análises sob uma perspectiva anual, visto que ocorrerá lacunas. Nesse contexto, o Apêndice B apresenta as 33 variáveis nesta situação.

<table>
<tr><th>2008</th><th>2009</th><th>2010</th><th>2011</th><th>2012</th><th>2013</th><th>2014</th><th>2015</th><th>2016</th><th>2017</th><th>2018</th><th>2019</th><th>2020</th></tr>
<tr><td colspan="13">ano_cmpt, car_int, cbor, cep, cgc_hosp, cid_asso, cid_morte, cid_notif, cnaer, cnes, cnpj_mant, cobranca, cod_idade, complex, contracep1, contracep2, cpf_aut, diag_princ, diag_secun, diar_acom, dias_perm, dt_inter, dt_saida, espec, etnia, faec_tp, financ, gestao, gestor_cod, gestor_cpf, gestor_dt, gestor_tp, gestrisco, homonimo, idade, ident, ind_vdrl, infehosp, insc_pn, instru, marca_uti, mes_cmpt, morte, munic_mov, munic_res, n_aih, nacional, nasc, natureza, num_filhos, num_proc, proc_rea, proc_solic, qt_diarias, raca_cor, regct, remessa, rubrica, seq_aih5, sequencia, sexo, tot_pt_sp, uf_zi, us_tot, uti_int_al, uti_int_an, uti_int_in, uti_int_to, uti_mes_al, uti_mes_an, uti_mes_in, uti_mes_to, val_acomp, val_obsang, val_ortp, val_ped1ac, val_rn, val_sadt, val_sadtsr, val_sangue, val_sh, val_sp, val_tot, val_transp, val_uti, vincprev</td></tr>
<tr><td colspan="4"></td><td colspan="9">aud_just, nat_jur, sis_just, val_sh_fed, val_sh_ges, val_sp_fed, val_sp_ges</td></tr>
<tr><td colspan="5"></td><td colspan="8">marca_uci, val_uci</td></tr>
<tr><td colspan="6"></td><td colspan="7">diagsec1, diagsec2, diagsec3, diagsec4, diagsec5, diagsec6, diagsec7, diagsec8, diagsec9, tpdisec1, tpdisec2, tpdisec3, tpdisec4, tpdisec5, tpdisec6, tpdisec7, tpdisec8, tpdisec9</td></tr>
</table>

Tabela 1: variáveis existentes por período.

Variável	Mudança	Período de ocorrência
aud_ust	Aumentou de tamanho: passou de um tamanho máximo de 50 para 68	2013
	Diminuiu de tamanho: passou de um tamanho máximo de 68 para 65	2015
	Diminuiu de tamanho: passou de um tamanho máximo de 68 para 59	2015
insc_n	Aumentou de tamanho: passou de um tamanho máximo de 10 para 12	2012
sis_ust	Aumentou de tamanho: passou de um tamanho máximo de 50 para 53	2013
	Aumentou de tamanho: passou de um tamanho máximo de 53 para 59	2014
	Diminuiu de tamanho: passou de um tamanho máximo de 59 para 50	2015
	Aumentou de tamanho: passou de um tamanho máximo de 50 para 54	2017
	Diminuiu de tamanho: passou de um tamanho máximo de 54 para 50	2018
val_h	Aumentou de tamanho: passou de um tamanho máximo de 8 para 9	2009
val_h_ed	Apenas valores [R$] 0	2011
val_h_es	Apenas valores [R$] 0	2011:2014
val_p_ed	Apenas valores [R$] 0	2011
val_p_es	Apenas valores [R$] 0	2011:2014

Tabela 2: mudanças ocorridas nas variáveis por período.

Todas as mudanças observadas neste estudo, descritas na Tabela 2 e no Apêndice B, são referentes ao tamanho máximo de caracteres permitido para os re-

gistros. Os valores *NA* são referentes a presença de valores faltantes/não preenchidos especificamente no período mencionado. Por exemplo, observa-se pelo Apêndice B que a variável *aud_just*, acrescentada na base de dados a partir do ano 2011, contém todos os registros do ano 2011 *NA*. Este fato, de **algumas variáveis apresentarem somente valores *NA* no primeiro ano de sua implantação na base**, se repete também para a maioria das variáveis que foram inseridas na base a partir de 2011, a saber: *aud_just*, *diagsec1*, *diagsec2*, *diagsec3*, *diagsec4*, *diagsec5*, *diagsec6*, *diagsec7*, *sis_just*, *tpdisec1*, *tpdisec2*, *tpdisec3*, *tpdisec4*, *tpdisec5*, *tpdisec6*, *tpdisec7*, *tpdisec8*, *tpdisec9*.

Destaca-se também que todas as variáveis que foram inseridas em 2014 apresentaram algum tipo de mudança de domínio ou problema, visto que as únicas duas não destacadas na Tabela 1, *diagsec8* e *diagsec9* possuem apenas registros não disponíveis.

Em virtude dos problemas destacados na Seção 4 e nas Tabelas 1 e 2, conclui-se que **não há nenhum ano sem apresentar qualquer tipo de problema relacionado a disponibilidade dos dados e ocorrência de alterações nas variáveis**.

6 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 são apresentadas no Apêndice A. Resultados numéricos dos testes de completude, conformidade e acurácia são exibidos no Apêndice C, onde estão organizados em três tabelas: resultado geral, resultado agregado por ano e resultado agregado por Unidade Federativa. O Apêndice D apresenta uma tabela contendo os registros mais e menos frequentes para as variáveis que não estão totalmente conformes. Já o Apêndice E 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 F. Os resultados agregados por ano foram obtidos em relação ao nome do arquivo.

6.1 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 au-

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

sente em linguagem R, e *NULL*, que representa objetos nulos. Ressalta-se, ainda, que houveram ajustes no total das variáveis, visto que 27 não encontram-se em todo o período representado pela base, conforme discutido na Seção 5.

No que tange a distribuição temporal dos resultados dos testes de completude apresentados no Gráfico 3, percebe-se que houve grande queda em 2014, passando de 86,96% em 2013 para 73,11%. Em seguida, o preenchimento dos registros voltou a aumentar, porém sem ultrapassar o que já era encontrado no início do período representado: 2008, com 85,39%.

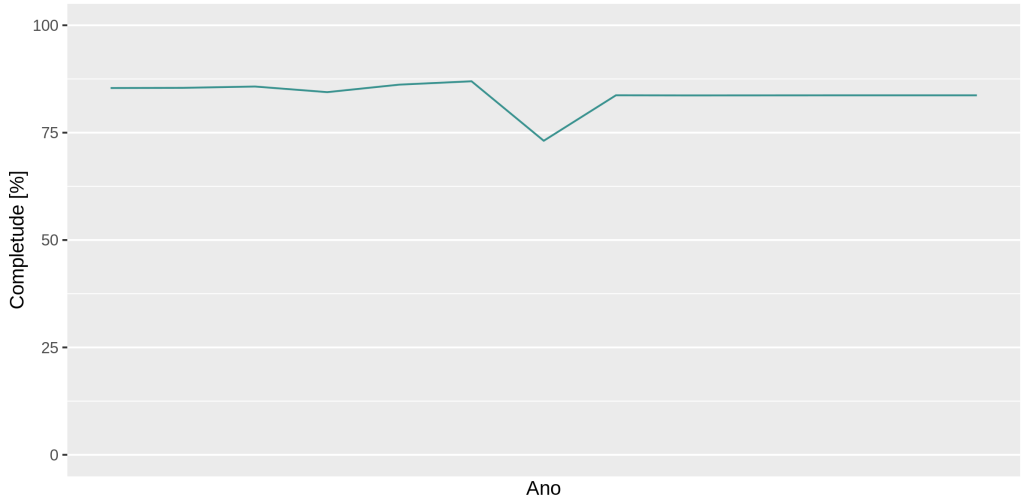

Gráfico 3: distribuição temporal da completude agregada por Unidade Federativa.

Já a respeito dos resultados distribuídos espacialmente, têm-se o Gráfico 4, onde nota-se que o preenchimento dos registros distribui-se em torno de 85% para todas as Unidades Federativas. Distrito Federal (86,55%) e Tocantins (86,51%) apresentaram as maiores taxas de preenchimento, enquanto Paraíba (85,07%), Pará (85,13%) e Goiás (85,16%) apresentaram as menores.

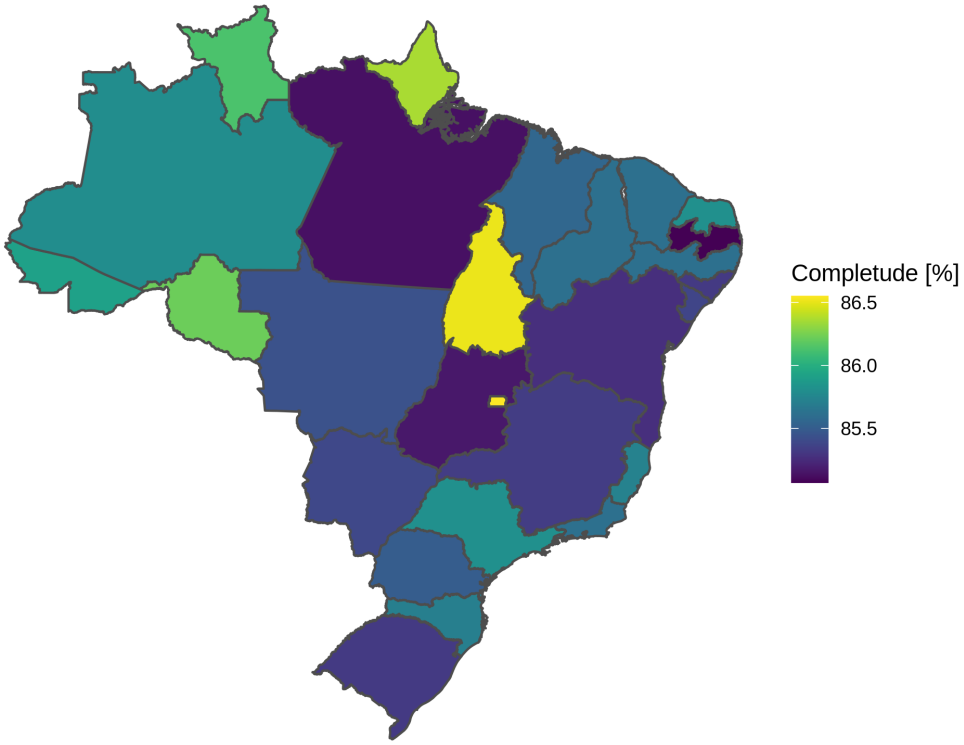

Gráfico 4: distribuição espacial da completude por Unidade Federativa agregada no tempo.

No geral, os resultados de completude das variáveis estão distribuídas pelas categorias definidas na Seção 3 segundo o Gráfico 5. O resultado percentual por variável está descrito no Apêndice C. O cômputo da média ponderada dos resultados obtidos é de **84,57%**, ou seja, a **completude é ótima**.

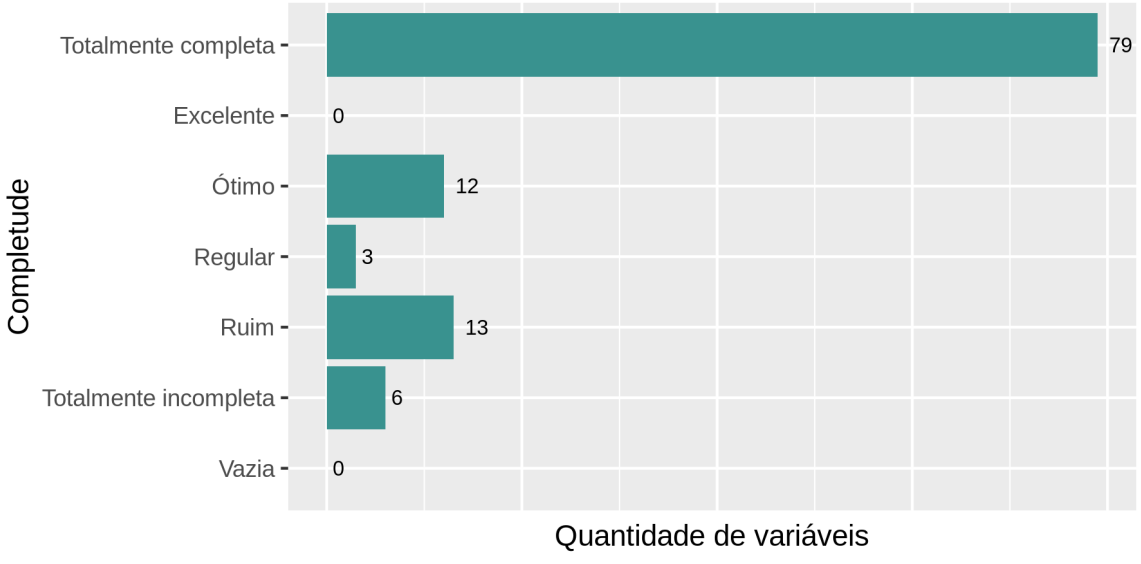

Gráfico 5: distribuição dos resultados de completude.

6.2 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 e presença de variantes definidos. O resultado percentual por variável está descrito no Apêndice C.

A Tabela 3 apresenta os resultados encontrados. Cinco tipos de erros referem-se ao tamanho dos registros, maior que o estipulado pelo dicionário de dados, e três referem-se a valores não definidos no domínio finito da variável.

Variável	**Falha identificada**
aud_just	2,90% ultrapassa os 50 caracteres definidos
cbor	100% possuem 6 dígitos, contrastando aos 3 dígitos definidos
gestor_cod	93,00% possui 5 dígitos, contrastando aos 3 dígitos definidos
gestor_cpf	95,10% possuem 15 dígitos, contrastando aos 11 dígitos (tamanho de CPF) definidos
instru	6 registros inconformes (valores *5*, *6* e *8*)
marca_uci	88 registros inconformes (valor *88*)
regct	0,24% inconformes (valores *7112* e *7113*)
sis_just	5 registros ultrapassam os 50 caracteres definidos

Tabela 3: registros inconformes

Cita-se, também 19 variáveis contendo a descrição *Zerado* segundo o dicionário adotado (Apêndice A). Em 17 delas, *rubrica*, *tot_pt_sp*, *uti_int_al*, *uti_int_an*, *uti_int_in*, *uti_mes_al*, *uti_mes_an*, *uti_mes_in*, *val_acomp*, *val_obsang*, *val_ortp*, *val_ped1ac*, *val_rn*, *val_sadt*, *val_sadtsr*, *val_sangue* e *val_transp*, há apenas a ocorrência do valor *0* (conforme e não acurado); e nas outras duas restantes, *cpf_aut* e *num_proc*, há apenas registros nulos.

Ao que diz respeito aos resultados dos testes de conformidade agregados por ano representados pelo Gráfico 6, houve ligeira queda com relação ao avanço do período, uma vez que o pico é atingido em 2008, de 98,13%. No restante dos anos, todos apresentam resultados bem próximos de 96% e 97%. De maneira semelhante, os resultados distribuídos espacialmente, Gráfico 4, também se apresentaram bem próximos de 96% e 97%.

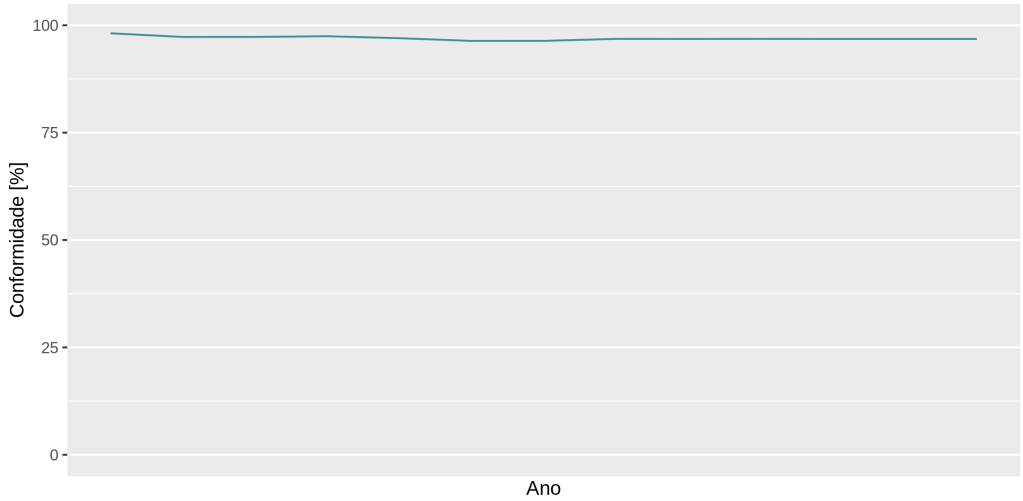

Gráfico 6: distribuição temporal da conformidade agregada por Unidade Federativa.

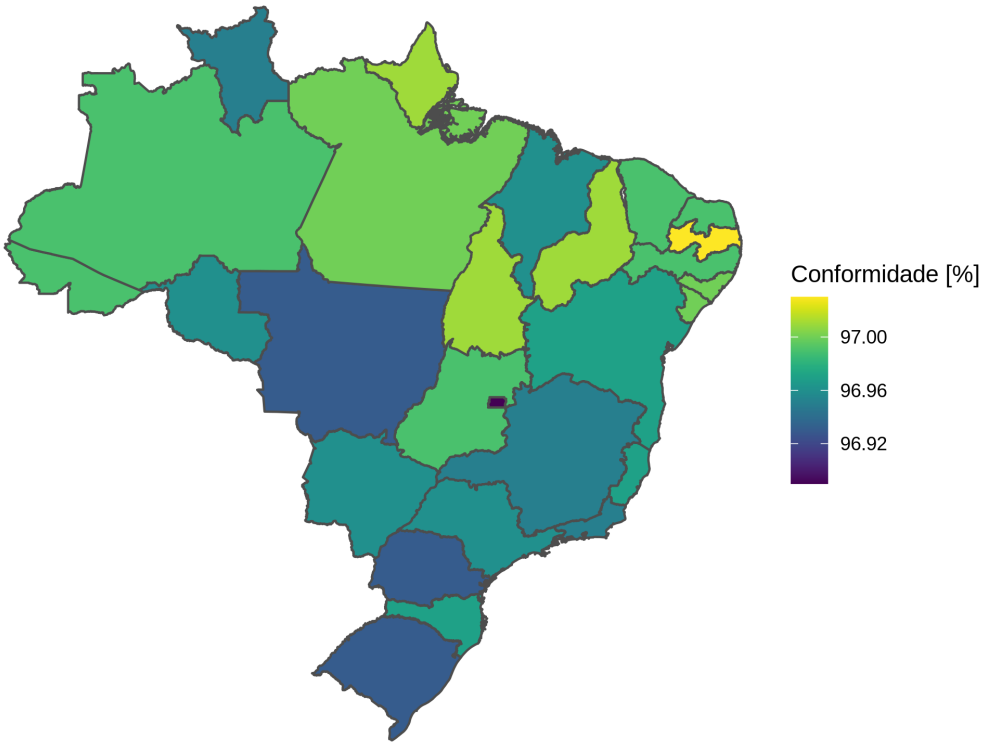

Gráfico 7: distribuição espacial da conformidade por Unidade Federativa agregada no tempo.

No geral, os resultados de conformidade das variáveis estão distribuídas pelas categorias definidas na Seção 3 segundo o Gráfico 8. O resultado percentual por variável está descrito no Apêndice C. O cômputo da média ponderada dos resultados obtidos é de **97,00%**, ou seja, a **conformidade é excelente**.

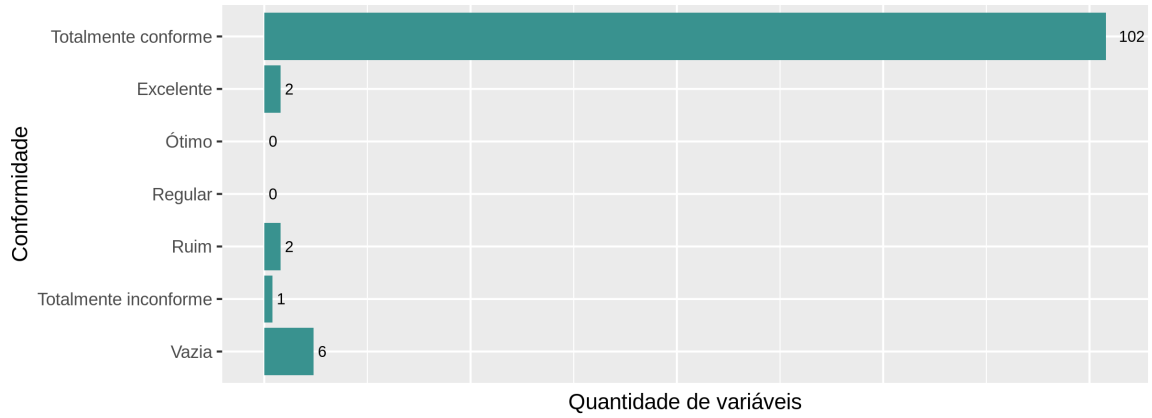

Gráfico 8: distribuição dos resultados de conformidade.

6.3 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).

A respeito de registros representando datas, apenas para a variável *nasc* verificou-se se existiam datas superiores ao mês e ano referência da base de dados analisada, isto é, último dia representado pela tabela de dados. Para as demais, verificou-se se as datas pertencem ao mês e ano referência da base analisada. É interessante ressaltar que na variável *nasc* há 2438 registros cujo ano de nascimento é menor que 1908, ou seja, há 2438 pacientes com mais de 100 anos. O mínimo encontrado nesta base é datado de 10 de outubro de 1887, ou seja, um paciente com no mínimo 121 anos desde o lançamento da informação no sistema.

Acerca dos demais registros, pode-se mencionar a ocorrência de valores indicando a ausência de informações ou que foram ignorados, além de sequências finitas do numeral zero. 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. Por exemplo, na variável *uf_zi*, que representa o código do município gestor, 42,6% dos registros referem-se a sequências de apenas zeros. Na variável *cnaer*, que representa o código de aci-

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

dente de trabalho, 99,99% dos registros também possuem apenas o numeral zero. Acerca dos registros sem informação, cita-se a variável *raca_cor*, representando a raça/cor do paciente. Nesta, 29,72% refere-se ao valor *99*, que de acordo com o dicionário de dados, representa *sem informação*.

Em relação as várias quantitativas, o Apêndice E expõe uma tabela contendo valores atípicos, os quais implicam, tipicamente, em prejuízos a interpretação dos resultados dos testes estatísticos aplicados. Neste, observa-se valores incomuns para todas as variáveis analisadas, como por exemplo pacientes com 121 anos de idade. Nesse mesmo contexto, cita-se a presença de 1 registro na variável *qt_diarias* contendo o valor negativo *-15*.

Acerca dos testes de acurácia no que diz respeito aos resultados distribuídos por ano, é possível observar através do Gráfico 9, que a acurácia dos registros claramente piorou com o decorrer do tempo. Partindo de 62,42% em 2008, ocorreram pequenos acréscimos até atingir o máximo de 65,88% em 2014, ano este que possui a menor completude (Subseção 6.1). Em seguida, sucederam significativas quedas, até atingir o mínimo em 2020, de 54,75%.

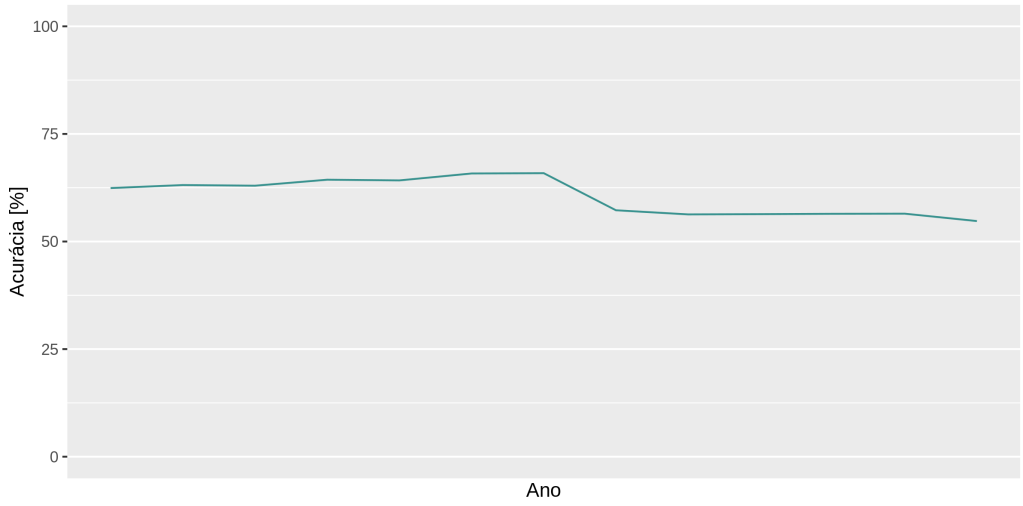

Gráfico 9: distribuição temporal da acurácia agregada por Unidade Federativa.

Já sobre a distribuição espacial, apresentada pelo Gráfico 10, têm-se que o mínimo, 57,92% é apresentado pelo Distrito Federal, estado este que possui melhor completude dos registros (Subseção 6.1). O máximo é encontrado pelo estado da Paraíba, de 61,33%. Novamente, entra em contraste ao fato de este estado possuir a menor completude dos registros.

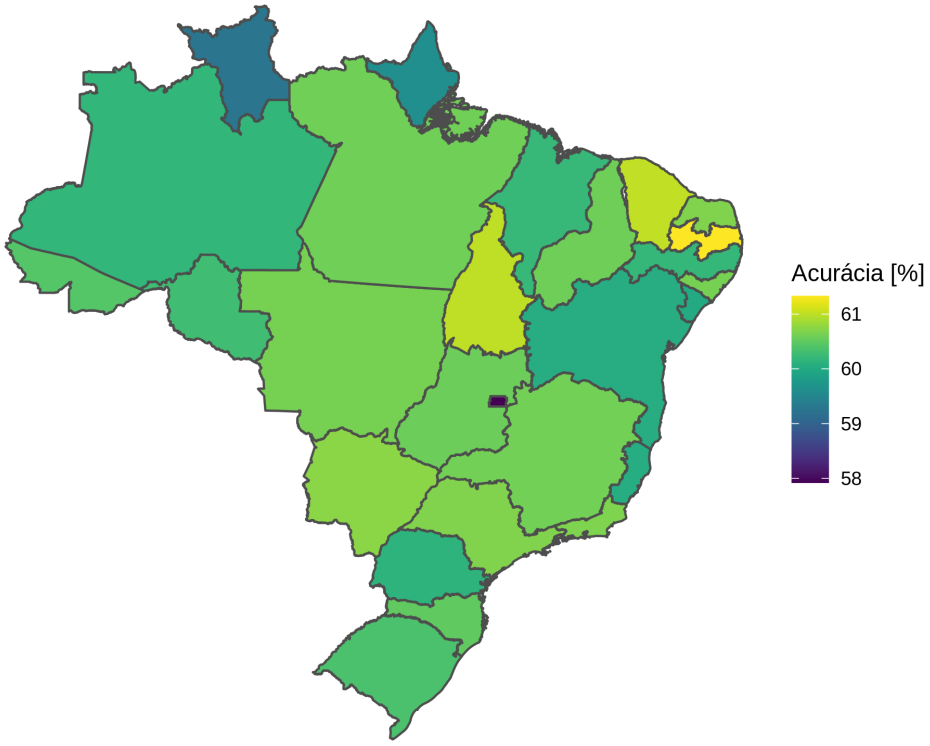

Gráfico 10: distribuição espacial da acurácia por Unidade Federativa agregada no tempo.

No geral, os resultados de completude das variáveis estão distribuídas pelas categorias definidas na Seção 3 segundo o Gráfico 5. O resultado percentual por variável está descrito no Apêndice C. O cômputo da média ponderada, considerando todos os testes realizados, é **74,87%**, ou seja, a **acurácia é regular** para a base analisada.

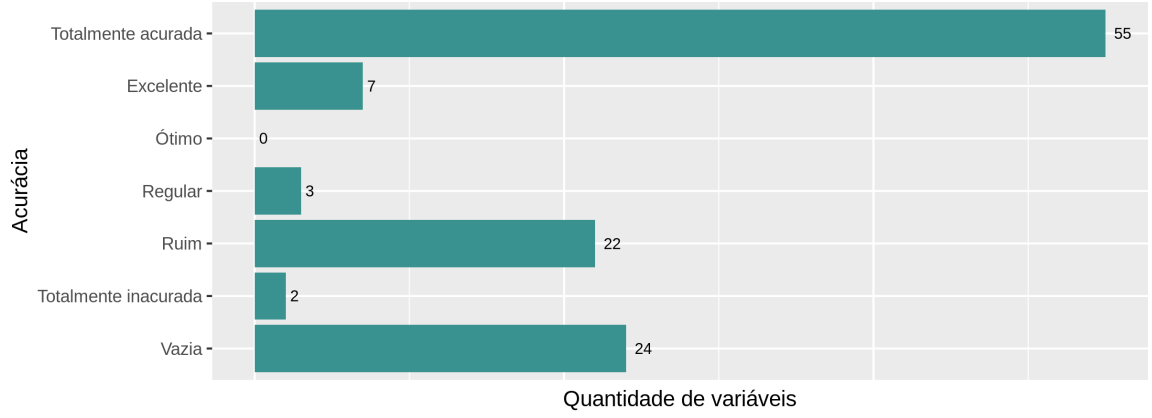

Gráfico 11: distribuição dos resultados de acurácia.

6.4 Consistência

Os resultados descritos na Tabela 4 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 realizados, bem como os resultados, se encontram no Apêndice F, onde também é destacada a quantidade de inconsistências por teste para cada ano.

Teste realizado	**Falhas [partes por mil]**
gentrisco == 0 & sexo == 1	1,607
insc_pn !=0 & sexo == 1	0,183
nasc >dt_saida	0,000
nasc >dt_inter	0,000
dt_inter >dt_saida	0,000

Tabela 4: resultados de consistência.

Sobre a distribuição temporal dos resultados de consistência, têm-se o Gráfico 12, onde nota-se que o ano com maior quantidade de registros inconsistentes foi 2008, com 4,136 partes por mil. Após 2008 os registros inconsistentes diminuíram consideravelmente.

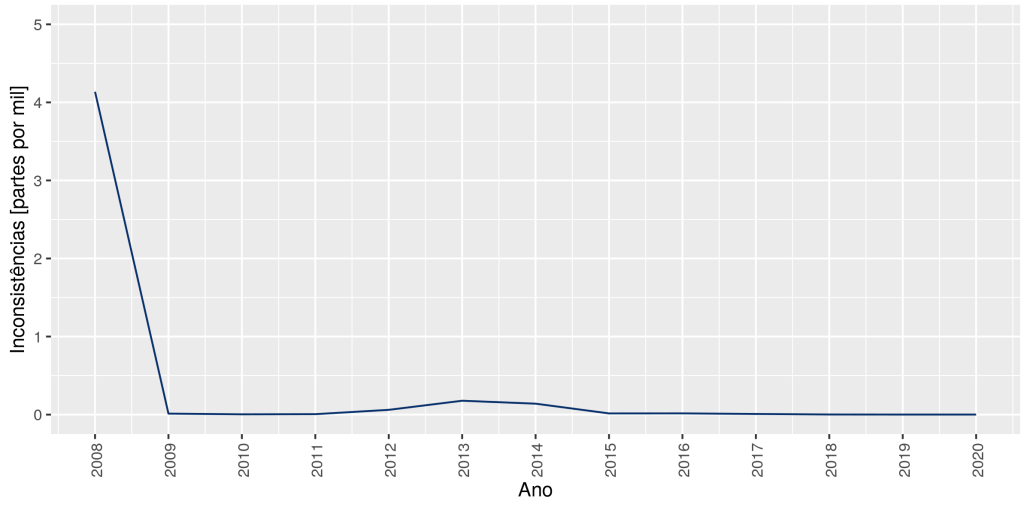

Gráfico 12: distribuição temporal de inconsistências agregadas por Unidade Federativa.

De forma análoga têm-se o Gráfico 13, onde é apresentada a distribuição espacial dos resultados de consistência. Observa-se que o estado do Paraná registrou o maior valor de inconsistência em comparação aos demais, com 0,846 inconsis-

tências por mil, enquanto o estado do Amapá registrou 0,000 inconsistências por mil.

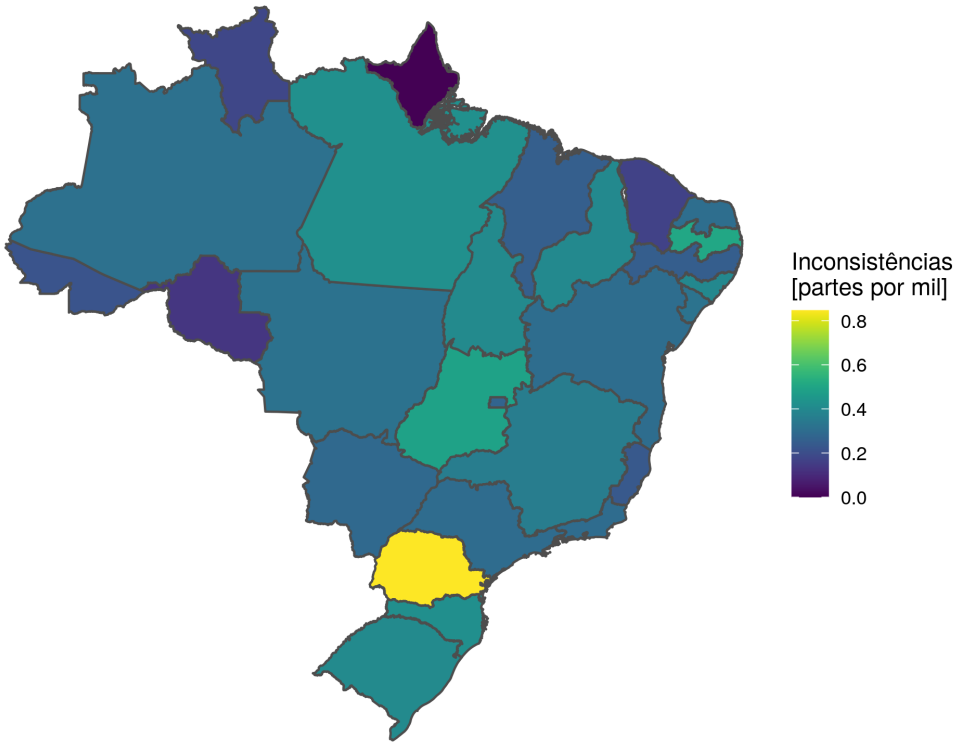

Gráfico 13: distribuição espacial das inconsistências por Unidade Federativa agregadas por ano.

No geral, a média ponderada dos resultados dos testes de consistência é **99,96%**, ou seja, a **consistência é excelente** para a base analisada.

6.5 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 5. O primeiro e único teste, *T1*, refere-se ao período de internação do paciente.

Teste	Variável inicial	Variável final	Mediana	Máx.	Min.
T1	*dt_inter*	*dt_saida*	3	4473	0

Tabela 5: resultados de temporalidade.

Buscando avaliar os resultados em uma escala temporal, o Gráfico 14 apresenta a evolução das medianas em relação ao ano representado, enquanto o Gráfico 15 apresenta os resultados em escala por unidade federativa. Nota-se que

houve um resultado uniforme da mediana em relação aos anos, enquanto os resultados relacionados as unidades federativas registram duas medianas: 2 e 3.

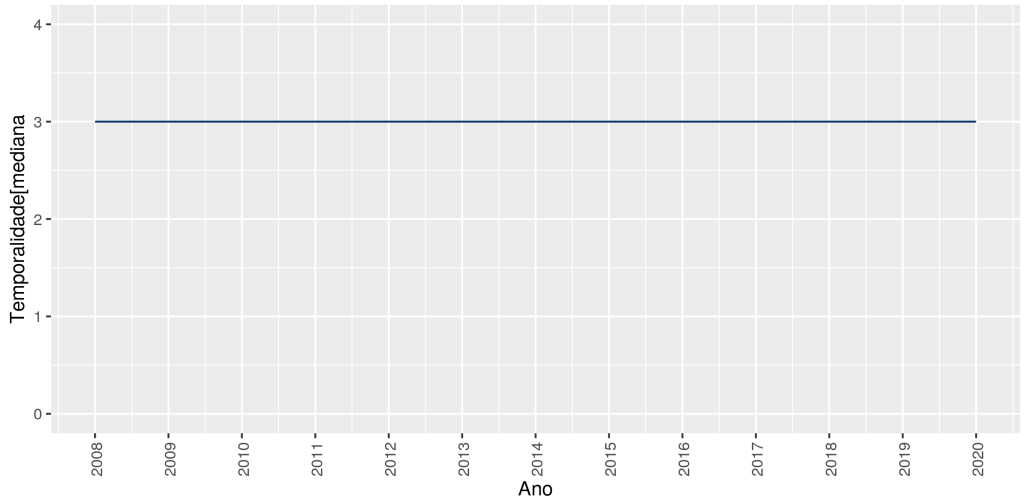

Gráfico 14: distribuição temporal das medianas de temporalidade agregadas por Unidade Federativa.

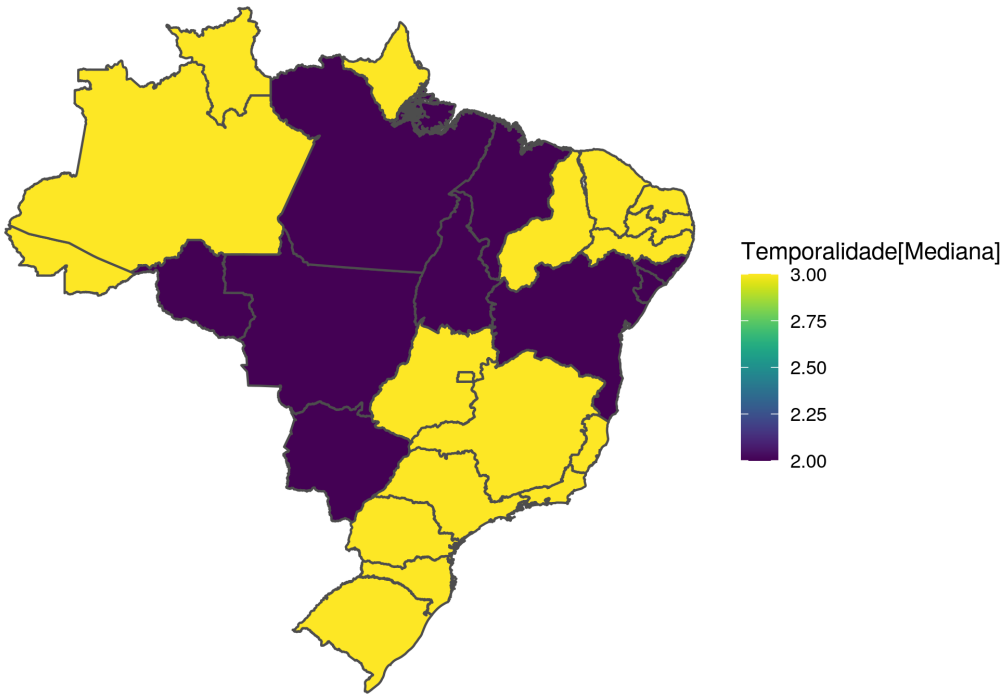

Gráfico 15: distribuição espacial das medianas de temporalidade por Unidade Federativa agregadas por ano.

7 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.

Avaliando a disponibilidade dos dados, ressalta-se que a falta de informações de determinados estados e períodos reflete em problemas na análise aqui realizada, visto que não irá representar a abrangência total dos registros. Em relação ao aumento das variáveis, isto é expressado pelo interesse de uma melhora do atendimento da demanda por informações provenientes dos órgãos de saúde pública. Algumas variáveis, entretanto, apresentaram problemas e merecem ser analisadas, com o intuito de incrementar a qualidade da informação dos dados e incentivar o preenchimento correto pelos profissionais de saúde envolvidos no processo. Sobre estas alterações numa escala temporal, conclui-se que não há nenhum ano sem apresentar qualquer tipo de deficiência.

Analisado os resultados obtidos pela métrica de completude, observa-se que, embora haja bons resultados, deve-se mencionar a grande ocorrência de sequências finitas do numeral zero, o que pode representar um problema visto que estará de acordo ao tamanho estabelecido pelo dicionário de dados, porém sem representar informação alguma. Quanto à distribuição temporal, houveram variações no período representado, da mesma forma na distribuição estadual, onde Distrito Federal e Paraíba apresentaram as maiores e menores taxas de preenchimento dos registros, respectivamente.

Sobre os resultados dos testes de conformidade, a maior quantidade numérica de falhas referem-se ao tamanho da representação, maior que o estabelecido pelo dicionário de dados adotado. Acerca dos resultados distribuídos por ano, houve ligeira queda com relação ao avanço do período, enquanto a distribuição estadual apresentou constância.

A respeito dos resultados de acurácia, houveram significativas falhas em todos os testes realizados. Dentre estes, destaca-se, novamente, a grande ocorrência de sequências finitas do numeral zero, além de valores discrepantes nas variáveis numéricas, como por exemplo pacientes com no mínimo 121 anos. Sobre a distribuição temporal a mesma piorou com o decorrer do tempo, o que leva a necessidade de medidas corretivas, Acerca da distribuição estadual, observou-se o inversão dos resultados de completude nos estados cujos registros estão mais e menor acurados, Paraíba e Distrito Federal. Ainda deve-se mencionar a ocorrência de registros ignorados e sem informação, um problema por não representar informação alguma.

Com relação aos resultados de consistência, conclui-se que o ano de 2008 registrou maior quantidade de inconsistências e os demais anos registraram menores quantidades de inconsistências. De forma geral, a base avaliada é bastante consistente, visto que foram realizados 5 testes de inconsistência e em mais metade não houveram ocorrência de falhas.

Sobre os resultados de temporalidade, foi possível observar variações nos testes apenas em relação as unidades federativas. Os valores máximos 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 84,57%, de conformidade é 97,00%, de acurácia é 74,87%, de consistência é 99,96%. Realizando o produto destes resultados, obtêm-se **61,40%**, caracterizando a **base de dados como regular**.

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 DATASUS. *SIHSUS - Sistema de Informações Hospitalares do SUS*. 2020. Online; acessado em 20 de maio de 2020. Disponível em: <http://datasus1.saude.gov.br/sistemas-e-aplicativos/hospitalares/sihsus>.

4 IBGE - Instituto Brasileiro de Geografia e Estatística. *Sistema de Informações Hospitalares do SUS - SIH/SUS*. 2020. Online; acessado em 20 de maio de 2020. Disponível em: <https://ces.ibge.gov.br/base-de-dados/metadados/ministerio-da-saude/sistema-de-informacoes-hospitalares-do-sus-sih-sus.html>.

5 CERQUEIRA, D. R. C. et al. *Uma análise da base de dados do Sistema de Informação Hospitalar entre 2001 e 2018: dicionário dinâmico, disponibilidade dos dados e aspectos metodológicos para a produção de indicadores sobre violência*. Rio de Janeiro, 2019.

6 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 / Observações	Tam.	Valores válidos
ano_cmpt	Ano de processamento da AIH, no formato aaaa.	4	
aud_just	Justificativa do auditor para aceitação da AIH sem o número do Cartão Nacional de Saúde.	50	
car_int	Caráter da internação.	2	01: Eletivo, 02: Urgência, 03: Acidente no local trabalho ou a serv da empresa, 04: Acidente no trajeto para o trabalho, 05: Outros tipo de acidente de trânsito, 06: Out tp lesões e envenen por agent quím físicos
cbor	Ocupação do paciente, segundo a Classificação Brasileira de Ocupações – CBO.	3	
cep	CEP do paciente.	8	
cgc_hosp	CNPJ do Estabelecimento.	14	
cid_asso	CID causa.	4	
cid_morte	CID da morte.	4	
cid_notif	CID de Notificação.	4	
cnaer	Código de acidente de trabalho.	3	
cnes	Código CNES do hospital.	7	
cnpj_mant	CNPJ da mantenedora.	14	
cobranca	Motivo de Saída/Permanência	2	
cod_idade	Unidade de medida da idade.	1	
complex	Complexidade.	2	00-99: Não se aplica, 01: Atenção Básica, 02: Média complexidade, 03: Alta complexidade
contracep1	Tipo de contraceptivo utilizado.	2	00-99: Ignorado/não se aplica, 01: LAM, 02: Ogino Knaus, 03: Temperatura basal, 04: Billings, 05: Cinto térmico, 06: DIU, 07: Diafragma, 08: Preservativo, 09: Espermicida, 10: Hormônio oral, 11: Hormônio injetável, 12: Coito interrompido

Variável	Descrição / Observações	Tam.	Valores válidos
contracep2	Segundo tipo de contraceptivo utilizado.	2	00-99: Ignorado/não se aplica, 01: LAM, 02: Ogino Knaus, 03: Temperatura basal, 04: Billings, 05: Cinto térmico, 06: DIU, 07: Diafragma, 08: Preservativo, 09: Espermicida, 10: Hormônio oral, 11: Hormônio injetável, 12: Coito interrompido
cpf_aut	Zerado	11	
diag_princ	Código do diagnóstico principal (CID10).	4	
diag_secun	Código do diagnostico secundário (CID10). Preenchido com zeros a partir de 201501.	4	
diagsec1	Diagnóstico secundário1.	4	
diagsec2	Diagnóstico secundário 2.	4	
diagsec3	Diagnóstico secundário 3.	4	
diagsec4	Diagnóstico secundário 4.	4	
diagsec5	Diagnóstico secundário 5.	4	
diagsec6	Diagnóstico secundário 6.	4	
diagsec7	Diagnóstico secundário 7.	4	
diagsec8	Diagnóstico secundário 8.	4	
diagsec9	Diagnóstico secundário 9.	4	
diar_acom	Quantidade de diárias de acompanhante.	3	
dias_perm	Dias de Permanência.	5	
dt_inter		8	
dt_saida	Data de saída, no formato aaaammdd.	8	
espec	Especialidade do Leito	2	
etnia	Etnia do paciente, se raça cor for indígena.	4	
faec_tp	Subtipo de financiamento FAEC.	6	
filename		255	
financ	Tipo de financiamento.	2	00-99: Não discriminado, 01: Atenção Básica (PAB), 02: Assistência Farmacêutica, 04: Fundo de Ações Estratégicas e Compensações FAEC, 05: Incentivo – MAC, 06: Média e Alta Complexidade (MAC), 07: Vigilância em Saúde

Variável	Descrição / Observações	Tam.	Valores válidos
gestao		1	0: Estadual, 2: Estadual plena, 1: Municipal plena assist, 3-9: Não definida
gestor_cod	Motivo de autorização da AIH pelo Gestor.	3	
gestor_cpf	Número do CPF do Gestor.	11	
gestor_dt	Data da autorização dada pelo Gestor (aaaammdd).	8	
gestor_tp	Tipo de gestor.	1	0: Estadual, 2: Estadual plena, 1: Municipal plena assist, 3-9: Não definida
gestrisco	Indicador se é gestante de risco.	1	1: Sim, 0: Não
homonimo	Indicador se o paciente da AIH é homônimo do paciente de outra AIH.	1	
idade	Idade.	2	
ident	Identificação do tipo da AIH.	1	0-9: Outras/ignorado, 1: Normal, 5: Longa permanência
ind_vdrl	Indica exame VDRL.	1	0: Sim, 1: Não
infehosp	Status de infecção hospitalar.	1	
insc_pn	Número da gestante no pré-natal.	12	
instru	Grau de instrução do paciente.	1	0-9: Ignorado/não se aplica, 1: Analfabeto, 2: 1 grau, 3: 2 grau, 4: 3 grau
marca_uci	Tipo de UCI utilizada pelo paciente.	2	00: Não utilizou UCI, 01: Unidade de cuidados intermed neonatal convencional, 02: Unidade de cuidados intermed neonatal canguru, 03: Unidade intermediária neonatal

Variável	Descrição / Observações	Tam.	Valores válidos
marca_uti	Indica qual o tipo de UTI utilizada pelo paciente.	2	00: Não utilizou UTI, 51: UTI adulto - tipo II COVID 19, 52: UTI pediátrica - tipo II COVID 19, 74: UTI adulto - tipo I, 75: UTI adulto - tipo II, 76: UTI adulto - tipo III, 77: UTI infantil - tipo I, 78: UTI infantil - tipo II, 79: UTI infantil - tipo III, 80: UTI neonatal - tipo I, 81: UTI neonatal - tipo II, 82: UTI neonatal - tipo III, 83: UTI de queimados, 85: UTI coronariana tipo II - UCO tipo II, 86: UTI coronariana tipo III - UCO tipo III, 99: UTI Doador, 01: Utilizou mais de um tipo de UTI
mes_cmpt	Mês de processamento da AIH, no formato mm.	2	01: Jan, 02: Fev, 03: Mar, 04: Abr, 05: Mai, 06: Jun, 07: Jul, 08: Ago, 09: Set, 10: Out, 11: Nov, 12: Dez, 00-13-99: Ign
morte	Indica Óbito	1	1: Com óbito, 0: Sem óbito
munic_mov	Município do Estabelecimento.	6	
munic_res	Município de Residência do Paciente	6	
n_aih	Número da AIH.	13	
nacional	Código da nacionalidade do paciente.	3	
nasc	Data de nascimento do paciente (aaaammdd).	8	
nat_jur	Natureza jurídica do Estabelecimento, conforme a Comissão Nacional de Classificação - CONCLA	4	
natureza	Natureza jurídica do hospital (com conteúdo até maio/12). Era utilizada a classificação de Regime e Natureza.:	2	00-99: Ignorado, 10: Próprio, 20: Contratado, 22: Contratado optante SIMPLES, 30: Federal, 31: Federal Verba Própria, 40: Estadual, 41: Estadual Verba Própria, 50: Municipal, 60: Filantrópico, 61: Filantrópico isento tributos e contr.sociais, 63: Filantrópico isento IR e contr.s/lucro líquido, 70: Universitário Ensino, 80: Sindicato, 90: Universitário Pesquisas, 91: Univ. Pesquisas isento tributos e contr.sociais, 93: Univ. Pesquisas isento IR e contr.s/lucro líquido, 94,92: Universitário de ensino e pesquisa privado
num_filhos	Número de filhos do paciente.	2	

Variável	Descrição / Observações	Tam.	Valores válidos
num_proc	Zerado	2	
proc_rea	Procedimento realizado.	10	
proc_solic	Procedimento solicitado.	10	
qt_diarias	Quantidade de diárias.	3	
raca_cor	Raça/Cor do paciente.	4	00-99: Sem informação, 01: Branca, 02: Preta, 03: Parda, 04: Amarela, 05: Indígena
regct	Regra contratual.	4	7101: S/crédito na média complex ambulatorial (exc FAEC), 7102: S/crédito na média complex hospitalar (exc FAEC), 7103: S/crédito na alta complex ambulatorial (exc FAEC), 7104: S/crédito na alta complex hospitalar (exc FAEC), 7105: S/crédito nos procedimentos financ FAEC, 7106: S/crédito total incluindo FAEC, 7107: S/crédito nas ações esp odonto (CEO I II III), 7108: S/crédito incentivo Saúde do Trabalhador exc FAEC, 7109: S/crédito total HU/MEC, 7110: S/crédito total Minist Saúde, 7111: S/crédito NASF exc FAEC, 0000: Sem regra contratual
remessa	Número da remessa.	21	
rubrica	Zerado	5	
seq_aih5	Sequencial de longa permanência (AIH tipo 5).	3	
sequencia	Sequencial da AIH na remessa.	9	
sexo	Sexo do paciente.	1	0-9: Ignorado, 1: Masculino, 2,3: Feminino
sis_just	Justificativa do estabelecimento para aceitação da AIH sem o número do Cartão Nacional de Saúde.	50	
tot_pt_sp	Zerado	6	
tpdisec1	Tipo de diagnóstico secundário 1.	1	
tpdisec2	Tipo de diagnóstico secundário 2.	1	
tpdisec3	Tipo de diagnóstico secundário 3.	1	
tpdisec4	Tipo de diagnóstico secundário 4.	1	
tpdisec5	Tipo de diagnóstico secundário 5.	1	

Variável	Descrição / Observações	Tam.	Valores válidos
tpdisec6	Tipo de diagnóstico secundário 6.	1	
tpdisec7	Tipo de diagnóstico secundário 7.	1	
tpdisec8	Tipo de diagnóstico secundário 8.	1	
tpdisec9	Tipo de diagnóstico secundário 9.	1	
uf_zi	Município Gestor.	6	
us_tot	Valor total, em dólar.	10	
uti_int_al	Zerado	2	
uti_int_an	Zerado	2	
uti_int_in	Zerado	2	
uti_int_to	Quantidade de diárias em unidade intermediaria.	3	
uti_mes_al	Zerado	2	
uti_mes_an	Zerado	2	
uti_mes_in	Zerado	2	
uti_mes_to	Quantidade de dias de UTI no mês.	3	
val_acomp	Zerado	13	
val_obsang	Zerado	11	
val_ortp	Zerado	13	
val_ped1ac	Zerado	11	
val_rn	Zerado	13	
val_sadt	Zerado	13	
val_sadtsr	Zerado	11	
val_sangue	Zerado	13	
val_sh	Valor de serviços hospitalares.	13	
val_sh_fed	Valor do complemento federal de serviços hospitalares. Está incluído no valor total da AIH.	10	
val_sh_ges	Valor do complemento do gestor (estadual ou municipal) de serviços hospitalares. Está incluído no valor totalda AIH.	10	
val_sp	Valor de serviços profissionais.	13	
val_sp_fed	Valor do complemento federal de serviços profissionais. Está incluído no valor total da AIH.	10	

Variável	Descrição / Observações	Tam.	Valores válidos
val_sp_ges	Valor do complemento do gestor (estadual ou municipal) de serviços profissionais. Está incluído no valor total da AIH.	10	
val_tot	Valor total da AIH.	14	
val_transp	Zerado	13	
val_uci	Valor de UCI.	10	
val_uti	Valor de UTI.	8	
vincprev	Vínculo com a Previdência.	1	0-9: Não classificado, 1: Autônomo, 2: Desempregado, 3: Aposentado, 4: Não segurado, 5: Empregado, 6: Empregador

Apêndice B Completude por período

Variáveis contendo apenas valores não disponíveis (em cinza) por período.

Variável	**Ano**												
	2008	*2009*	*2010*	*2011*	*2012*	*2013*	*2014*	*2015*	*2016*	*2017*	*2018*	*2019*	*2020*
aud_just													
cpf_aut													
diagsec1													
diagsec2													
diagsec3													
diagsec4													
diagsec5													
diagsec6													
diagsec7													
diagsec8													
diagsec9													
etnia													
gestor_cod													
gestor_dt													
infehosp													
marca_uci													
nat_jur													
num_proc													
sis_just													
tpdisec1													
tpdisec2													
tpdisec3													
tpdisec4													
tpdisec5													
tpdisec6													
tpdisec7													
tpdisec8													

Variável	**Ano**												
	2008	*2009*	*2010*	*2011*	*2012*	*2013*	*2014*	*2015*	*2016*	*2017*	*2018*	*2019*	*2020*
tpdisec9													
val_sh_fed													
val_sh_ges													
val_sp_fed													
val_sp_ges													
val_uci													

Apêndice C Resultados numéricos

C.1 Resultados gerais

Variável	Completude [%]	Conformidade [%]	Acurácia [%]
ano_cmpt	100,00	100,00	100,00
aud_just	1,31	97,09	99,80
car_int	100,00	100,00	100,00
cbor	100,00	0,00	NA
cep	100,00	100,00	100,00
cgc_hosp	76,56	100,00	100,00
cid_asso	44,12	100,00	1,41
cid_morte	45,52	100,00	4,44
cid_notif	0,81	100,00	100,00
cnaer	100,00	100,00	0,01
cnes	100,00	100,00	100,00
cnpj_mant	52,98	100,00	100,00
cobranca	100,00	100,00	100,00
cod_idade	100,00	100,00	100,00
complex	100,00	100,00	100,00
contracep1	100,00	100,00	0,82
contracep2	100,00	100,00	0,54
cpf_aut	0,00	NA	NA
diag_princ	100,00	100,00	100,00
diag_secun	51,00	100,00	14,71
diagsec1	16,09	100,00	100,00
diagsec2	1,54	100,00	100,00
diagsec3	0,55	100,00	100,00
diagsec4	0,21	100,00	100,00
diagsec5	0,08	100,00	100,00
diagsec6	0,03	100,00	100,00
diagsec7	0,01	100,00	100,00
diagsec8	0,00	NA	NA

Variável	**Completude [%]**	**Conformidade [%]**	**Acurácia [%]**
diagsec9	0,00	NA	NA
diar_acom	100,00	100,00	39,43
dias_perm	100,00	100,00	100,00
dt_inter	100,00	100,00	98,07
dt_saida	100,00	100,00	98,04
espec	100,00	100,00	100,00
etnia	77,95	100,00	0,21
faec_tp	2,63	100,00	100,00
financ	100,00	100,00	100,00
gestao	100,00	100,00	100,00
gestor_cod	67,75	7,00	5,73
gestor_cpf	100,00	4,90	11,59
gestor_dt	0,00	NA	NA
gestor_tp	100,00	100,00	100,00
gestrisco	100,00	100,00	100,00
homonimo	100,00	100,00	5,55
idade	100,00	100,00	98,69
ident	100,00	100,00	100,00
ind_vdrl	100,00	100,00	100,00
infehosp	0,00	NA	NA
insc_pn	100,00	100,00	4,17
instru	100,00	100,00	0,82
marca_uci	100,00	100,00	100,00
marca_uti	100,00	100,00	100,00
mes_cmpt	100,00	100,00	100,00
morte	100,00	100,00	100,00
munic_mov	100,00	100,00	99,18
munic_res	100,00	100,00	99,11
n_aih	100,00	100,00	100,00
nacional	100,00	100,00	100,00
nasc	100,00	100,00	99,95

Variável	**Completude [%]**	**Conformidade [%]**	**Acurácia [%]**
nat_jur	84,71	100,00	100,00
natureza	100,00	100,00	63,35
num_filhos	100,00	100,00	0,81
num_proc	0,00	NA	NA
proc_rea	100,00	100,00	100,00
proc_solic	100,00	100,00	100,00
qt_diarias	100,00	100,00	100,00
raca_cor	100,00	100,00	70,28
regct	100,00	99,76	100,00
remessa	100,00	100,00	100,00
rubrica	100,00	100,00	NA
seq_aih5	100,00	100,00	0,03
sequencia	100,00	100,00	100,00
sexo	100,00	100,00	100,00
sis_just	1,31	100,00	100,00
tot_pt_sp	100,00	100,00	NA
tpdisec1	84,23	100,00	19,10
tpdisec2	84,23	100,00	1,83
tpdisec3	84,23	100,00	0,65
tpdisec4	84,23	100,00	0,25
tpdisec5	84,23	100,00	0,09
tpdisec6	84,23	100,00	0,04
tpdisec7	84,23	100,00	0,01
tpdisec8	84,23	100,00	0,00
tpdisec9	84,23	100,00	0,00
uf_zi	100,00	100,00	57,40
us_tot	100,00	100,00	100,00
uti_int_al	100,00	100,00	NA
uti_int_an	100,00	100,00	NA
uti_int_in	100,00	100,00	NA
uti_int_to	100,00	100,00	100,00

Variável	Completude [%]	Conformidade [%]	Acurácia [%]
uti_mes_al	100,00	100,00	NA
uti_mes_an	100,00	100,00	NA
uti_mes_in	100,00	100,00	NA
uti_mes_to	100,00	100,00	100,00
val_acomp	100,00	100,00	NA
val_obsang	100,00	100,00	NA
val_ortp	100,00	100,00	NA
val_ped1ac	100,00	100,00	NA
val_rn	100,00	100,00	NA
val_sadt	100,00	100,00	NA
val_sadtsr	100,00	100,00	NA
val_sangue	100,00	100,00	NA
val_sh	100,00	100,00	100,00
val_sh_fed	100,00	100,00	100,00
val_sh_ges	100,00	100,00	100,00
val_sp	100,00	100,00	100,00
val_sp_fed	100,00	100,00	100,00
val_sp_ges	100,00	100,00	100,00
val_tot	100,00	100,00	100,00
val_transp	100,00	100,00	NA
val_uci	100,00	100,00	100,00
val_uti	100,00	100,00	100,00
vincprev	100,00	100,00	100,00

NA Apenas registros nulos.

C.2 Resultados por ano

Ano	Completude [%]	Conformidade [%]	Acurácia [%]
2008	85,39	98,13	62,42
2009	85,43	97,28	63,12
2010	85,74	97,29	62,97

Ano	Completude [%]	Conformidade [%]	Acurácia [%]
2011	84,43	97,45	64,36
2012	86,18	96,99	64,20
2013	86,96	96,36	65,81
2014	73,11	96,37	65,88
2015	83,70	96,83	57,26
2016	83,67	96,82	56,30
2017	83,68	96,83	56,36
2018	83,69	96,82	56,44
2019	83,69	96,82	56,47
2020	83,69	96,82	54,75

C.3 Resultados por Unidade Federativa

UF	Completude [%]	Conformidade [%]	Acurácia [%]
AC	85,91	96,99	60,43
AL	85,32	97,00	60,63
AM	85,79	96,99	60,16
AP	86,36	97,01	59,60
BA	85,27	96,97	60,05
CE	85,61	96,99	61,01
DF	86,55	96,89	57,92
ES	85,72	96,97	60,03
GO	85,16	96,99	60,57
MA	85,56	96,96	60,21
MG	85,33	96,95	60,61
MS	85,39	96,96	60,74
MT	85,44	96,93	60,63
PA	85,13	97,00	60,59
PB	85,07	97,03	61,33
PE	85,61	96,99	60,19
PI	85,61	97,01	60,58
PR	85,50	96,93	60,13

UF	**Completude [%]**	**Conformidade [%]**	**Acurácia [%]**
RJ	85,61	96,95	60,67
RN	85,81	96,99	60,69
RO	86,22	96,96	60,29
RR	86,13	96,95	59,24
RS	85,32	96,93	60,36
SC	85,70	96,97	60,50
SE	85,37	97,00	59,99
SP	85,81	96,96	60,69
TO	86,51	97,01	61,00

Apêndice D Registros mais e menos frequentes

Variável	**Registros mais frequentes**	**Registros menos frequentes**
aud_just	NA, PACIENTE NAO APRESENTOU CNS., SEM INFORMACAO, PACIENTE NAO TEM CNS, ADEQUANDO CADASTRO DO PACIENTE AO CADSUS.	CORRE<80>AO DO CNS EFETUADA COM SUCESSO, RN NAO TE CARTAO SUS, ADULTO COM ALTA MEDICA MELHORADA, PAC SEM CARTaO SUS, RN DE JOELMA ALVES DA SILVA COM DECLARA<a8><a8>O DE OBIT
cbor	000000, 225125, 515105, 252105, 225133	141405, 342115, 317210, 715120, 773505
gestor_cod	00000, NA, 000, 00007, 00134	169, 144, 110, 178, 049
gestor_cpf	000000000000000, 00000000000, 000012022594134, 000059470925068, 000010367578549	000090635485320, 000018697348668, 000008437006880, 000058909192615, 000005545055410
instru	0, 2, 3, 1, 4	4, 6, 8, 9, 5
marca_uci	00, NA, 01, 03, 02	NA, 01, 03, 02, 88
marca_uci	00, NA, 01, 03, 02	NA, 01, 03, 02, 88
regct	0000, 7102, 7109, 7106, 7104	7104, 7110, 7112, 7113, 7105
sis_just	NA, PACIENTE SEM CNS, PACIENTE NAO APRESENTOU CNS, URGENCIA, ATENDIMENTO DE EMERGENCIA CNS NAO OBTIDO	FALTADOCUMENTO, PACIENTE SEM CARTAO SUS INTERNET RUIM, PACIENTE COM DOCUMENTO INCOMPLETO, FOI INFORMADA A MAE PARA FAZER O CNS, RN COM CNS INVALIDO

Apêndice E Valores atípicos

Variável	Valores atípicos
dias_perm	10, 11, 12, 13, 14, 15, 16, 17, 18, 19, ...[3], 355, 356, 357, 358, 359, 360, 361, 362, 363, 364
gestor_dt	*
qt_diarias	10, 11, 12, 13, 14, 15, 16, 17, 18, 19, ..., 355, 357, 358, 359, 360, 361, 362, 363, 364, 365
seq_aih5	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99
sequencia	43104 124463, 43105 124464, 43106 124465, 43107 124466, 43108 124467, 43109 124468, 43110 124469, 43111 124475, 43112 124476, 43113 124477
us_tot	703.88, 703.89, 703.9, 703.91, 703.92, 703.93, 703.94, 703.95, 703.96, 703.97, ...[3], 80629.05, 82776.05, 83212.89, 84327.59, 86443.47, 88156.17, 89408.92, 89486.37, 97950.87, 98681.07
uti_int_al	*
uti_int_an	*
uti_int_in	*
uti_int_to	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, ...[3], 163, 170, 184, 192, 193, 205, 220, 228, 229, 245
uti_mes_al	*
uti_mes_an	*
uti_mes_in	*
uti_mes_to	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, ...[3], 292, 295, 298, 300, 302, 305, 306, 315, 334, 335
val_acomp	*
val_obsang	*
val_ortp	*
val_ped1ac	*
val_rn	*
val_sadt	*
val_sadtsr	*
val_sangue	*
val_sh	1235.22, 1235.23, 1235.24, 1235.25, 1235.26, 1235.27, 1235.28, 1235.29, 1235.3, 1235.31, ...[3], 189349.41, 190736.17, 190860.24, 191093.01, 191839.15, 192592.31, 200811.43, 201401.94, 204635.18, 207413.54

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

Variável	**Valores atípicos**
val_sh_fed	0.9, 0.95, 1.49, 1.56, 1.72, 1.77, 2, 2.04, 2.06, 2.23, ...[3], 4009.28, 4031.25, 4076.25, 4492.28, 4610.68, 5000, 5080.6, 5166.12, 5207.12, 7620.9
val_sh_ges	0.54 2581.56, 0.72 2583.06, 2.24 2591.82, 3.36 2678.86, 4.49 2716.76, 6.5 3344.86, 6.91 3591.27, 7.5 4000, 7.7 4009.28, 7.77 9409.32
val_sp	533.43 42217.87, 533.44 42537.19, 533.45 43519.87, 533.46 43868.37, 533.47 45413.9, 533.48 45777.25, 533.49 46661.39, 533.5 50700.71, 533.51 53479.19, 533.52 72642.67
val_sp_fed	0.37, 0.62, 0.73, 0.75, 0.76, 0.87, 0.92, 1.1, 1.24, 1.39, ...[3], 2694.46, 2765.44, 2917.26, 2920.26, 3317.72, 3484, 4614.89, 5000, 9547, 44769
val_sp_ges	1, 2.67, 3, 4, 4.25, 5, 6, 6.7, 8.01, ...[3], 1927.92, 1929.24, 1967.54, 2137.07, 2137.08, 2233.99, 2308.61, 2879.13, 2884.55, 3235.55
val_tot	1509.38, 1509.39, 1509.4, 1509.41, 1509.42, 1509.43, 1509.44, 1509.45, 1509.46, 1509.47, ...[3], 226698, 227992.77, 228225.54, 228408.85, 228971.68, 230264.99, 231734.02, 238534.47, 238808.2, 239929.94
val_transp	*
val_uci	136, 137.2, 150, 164.63, 171.49, 172.5, 180, 185.21, 187.5, 202.5, ...[3], 26342.4, 26460, 26479.6, 27180, 31050, 31281.6, 31418.8, 36900, 39600, 43170
val_uti	98, 117.59, 122.49, 132.29, 137.2, 139, 147, 152.9, 159.84, 164.63, ...[3], 152589, 152592, 154530.8, 155132.1, 160371.2, 169882.4, 178529.1, 190650.2, 194397.3, 203575.6

* Sem ocorrência de valores atípicos.

Apêndice F Testes de inconsistências

F.1 Testes realizados

- **Teste 1:** Se o indicador de gestante de risco for *0* (*Sim*) o sexo do paciente precisa diferente de *1* (*Masculino*).
- **Teste 2:** Se o numero da gestante no pré natal for diferente de *0*, logo existente, o sexo do paciente precisa diferente de *1* (*Masculino*).
- **Teste 3:** A data de nascimento do paciente não pode ser maior que a data de saída do paciente.
- **Teste 4:** A data de nascimento do paciente não pode ser maior que a data de internação do paciente
- **Teste 5:** A data de internação não pode ser maior que a data de saída do paciente.

F.2 Resultados obtidos

Ano	**Teste 1**	**Teste 2**	**Teste 3**	**Teste 4**	**Teste 5**
2008	229068	642	0	0	0
2009	0	736	0	0	0
2010	0	206	0	0	0
2011	0	331	0	0	0
2012	0	3497	0	0	0
2013	0	10233	0	0	0
2014	0	8127	0	0	0
2015	0	883	0	0	0
2016	0	906	0	0	0
2017	0	475	0	0	0
2018	1	83	0	0	0
2019	7	34	0	0	0
2020	1	8	0	0	0