UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE MEDICINA
PROGRAMA DE PÓS-GRADUAÇÃO EM EPIDEMIOLOGIA

TESE DE DOUTORADO

Pareamento de registros das grandes bases do SUS para permitir análises longitudinais de pacientes com câncer

Isaías Valente Prestes

Orientador: Prof. Dr. Bruce Bartholow Duncan

Porto Alegre, Agosto de 2017

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE MEDICINA
PROGRAMA DE PÓS-GRADUAÇÃO EM EPIDEMIOLOGIA

TESE DE DOUTORADO

Pareamento de registros das grandes bases do SUS para permitir análises longitudinais de pacientes com câncer

Isaías V. Prestes

Orientador: Prof. Dr. Bruce Bartholow Duncan

A apresentação desta tese é exigência do Programa de Pós-graduação em Epidemiologia, Universidade Federal do Rio Grande do Sul, para obtenção do título de Doutor.

Porto Alegre, Brasil.
2017

BANCA EXAMINADORA

Profa. Dra. Maria Paula Curado – A.C.Camargo Cancer Center / International Prevention Research Institute (iPRI).

Prof. Dr. Marcos Ennes Barreto – Universidade Federal da Bahia, Departamento de Ciência da Computação – UFBA.

Prof. Dr. Álvaro Vigo – Programa de Pós-graduação em Epidemiologia, Universidade Federal Rio Grande do Sul.

MENSAGEM

"Uncertainty, in the presence of vivid hopes and fears,
is painful, but must be endured if we wish to live
without the support of comforting fairy tales."

Bertrand Russell

AGRADECIMENTOS

Aos meus país, que para mim são simplesmente a origem de tudo.

Aos professores Bruce Bartholow Duncan e Maria Inês Schimidt, pelo que me ensinaram, pela paciência e o pelo esforço para lapidar o trabalho, mas, acima de tudo, pelo que criaram no País ao longo de suas carreiras – há anos promovendo o conhecimento e a vida de tantas pessoas.

Ao meu grande amigo Lenildo de Moura, sujeito que um dia apareceu na minha frente com 'problema nos dados', que me fez ir longe com esse problema e entender que o mundo é muito maior do que eu supunha.

I would like to express my appreciation to Dr. Timothy L. Lash for my academic and cultural experience in Emory University, introducing new ideas to me, guidance, and patience during this period.

Aos membros de nossa equipe, Gabriela Feiden, Silvana Roberta Lima e William Dartora pelo apoio ao trabalho ao longo desses anos – o "carregar o piano" do dia-a-dia sob sol ou chuva.

Ao Programa de Pós-Graduação em Epidemiologia da Faculdade de Medicina, que no meu mestrado e agora, oportunizou uma formação acadêmica de qualidade.

SUMÁRIO

ABREVIATURAS E SIGLAS

APAC - Autorização de Procedimentos de Alta Complexidade
BPA - Boletim de Produção Ambulatorial
BPA-C – Boletim de Produção Ambulatorial Consolidado
BPA-I – Boletim de Produção Ambulatorial Individualizado
CBO - Classificação Brasileira de Ocupações
CDC - Centers for Disease Control and Prevention
CNES - Cadastro Nacional de Estabelecimentos de Saúde
CNS - Cartão Nacional de Saúde
CPF - Cadastro de Pessoa Física
DANT - Doenças e Agravos Não Transmissíveis
DATASUS - Departamento de Informática do Sistema Único de Saúde
DBC - DataBase File compactado
DBF - DataBase File
DCNT - Doenças Crônicas Não Transmissíveis
GM - Gabinete do Ministro da Saúde
IARC - International Agency for Research on Cancer
IBGE - Instituto Brasileiro de Geografia e Estatística
INCA - Instituto Nacional de Câncer
MS - Ministério da Saúde
NCI - National Cancer Institute of United States of America
NHANES - National Health and Nutrition Examination Survey
NIH - National Health Institute
OMS - Organização Mundial de Saúde
PET-CT - Tomografia por emissão de pósitrons
PMP - Paciente por Milhão de População
PNAD - Pesquisa Nacional por Amostra de Domicílios
SAS – Secretaria de Atenção à Saúde
SAS/DRAC - Departamento de Regulação, Avaliação e Controle de Sistemas
SIA - Sistema de Informações Ambulatoriais
SIH - Sistema de Informações Hospitalares
SIM - Sistema de Informações sobre Mortalidade

SISCAN - Sistema de Informação do Câncer

SLANH - Sociedade Latino Americana de Nefrologia e Diálise Hipertensão

SUS - Sistema Único de Saúde

TRS - Terapia Renal Substitutiva

TRS - Terapia Renal Substitutiva

UE - União Europeia

UFRGS - Universidade Federal do Rio Grande do Sul

RESUMO

OBJETIVO: gerar uma base de dados de pacientes com câncer construída por meio de pareamento de registros das grandes bases de dados do SUS para permitir análise longitudinal do diagnóstico ao tratamento dos casos mais suscetíveis a tratamento, com ênfase em um conjunto de canceres potencialmente tratáveis e/ou frequentes.

MÉTODO: realizamos um estudo descritivo utilizando bases de dados do SUS com dados nominais, que foram disponibilizados pelo DATASUS. A organização da base de dados oncológicos iniciou com os dados da base de Autorização de Procedimentos de Alta Complexidade (SIA/APAC), do período de 2000 a 2012, para quimioterapia e radioterapia. Essas bases de APAC tiveram suas informações combinadas, formando uma única base oncológica. Os pacientes submetidos a tratamento de quimioterapia e/ou radioterapia receberam identificador único derivado da aplicação de pareamento probabilístico de registros. Essa identificação possibilitou a avaliação da trajetória do paciente ao longo do tratamento oncológico do SUS. Utilizamos o pareamento probabilístico de registros em outras duas etapas: 1) vinculação da base de químio-radioterapia com informações de hospitalização oriundas do Sistema de Informações Hospitalares do SUS (SIH/SUS); 2) vinculação de registros da base de oncologia com o Sistema de Informação sobre Mortalidade (SIM) – corrigindo informação sobre óbito, importante para futura análise de sobrevida. A base de dados de oncologia foi desenvolvida e controlada no programa de computador Statistical Analysis SAS®. Realizamos os pareamentos de registros no FRIL versão 2.1.5. Sobre esta base final, computamos estatísticas descritivas para os cânceres passíveis de prevenção, detecção precoce e tratamento definidos por Farmer et al. 2010. Calculamos as tendências das medidas epidemiológicas desses cânceres por meio do programa de computador JoinPoint Regression, versão 4.5.0.0.

RESULTADOS: construímos uma base de dados para oncologia, de âmbito nacional, orientada para análise epidemiológica, com informações de pacientes atendidos pelo SUS. Obtivemos estimativas de incidência de tratamentos e exploramos evolução do tempo entre diagnóstico e início de tratamento para o conjunto de canceres potencialmente tratáveis e/ou frequentes, no SUS. A análise epidemiológica do tratamento oncológico público no Brasil mostrou que o SUS ampliou o acesso ao

cuidado oncológico no âmbito de quimio e radioterapia nesse período. O tempo de espera entre diagnóstico e início do tratamento é longo e não se atenuou entre 2008 e 2012 consideravelmente.

CONCLUSÃO: o pareamento de dados se mostrou um processo efetivo para criação de uma base nacional de pacientes com câncer em tratamento de quimio- e/ou radioterapia no SUS, entre 2000 e 2012. A base de dados APAC oncologia concentra importante informação epidemiológica sobre tratamento do câncer no SUS. Dessa forma, ela apresenta grande potencial para ser usada no sentido de prover informação de serviço e políticas públicas. Além disso, representa uma importante ferramenta para o Ministério da Saúde desenvolver indicadores e monitorar o câncer no País – itens importantes assumidos no Plano de Ações Estratégicas para o Enfrentamento das Doenças Crônicas Não Transmissíveis no Brasil e a definição das metas globais para o enfrentamento dessas doenças até 2025.

ABSTRACT

OBJECTIVE: creation of a national database of cancer patients whose care was financed by the Brazilian national health system (SUS) through record linkage, in order to perform a longitudinal analysis of diagnosis through the treatment of cases most susceptible to treatment, with an emphasis on a set of potentially treatable cancers and / or frequent.

METHODS: the data was obtained from SUS databases provided by the Ministry of Health with nominal data permitting linkage. The creation of this cancer database begins with the database of High Complexity Procedures Authorization (SIA / APAC), period from 2000 to 2012, including chemotherapy and radiotherapy. Chemo- and radiotherapy database information has been combined into a single cancer database. A unique identifier derived from the application of probabilistic matching records was set to each patient receiving these therapies. This identification process enabled an evaluation of the patient trajectory through oncological diagnosis and treatment financed by SUS. Probabilistic record linkage was also used in other two instances: 1) linking the APAC chemo-radiotherapy (oncology) database with hospitalization information from the SUS Hospital Information System (SIH / SUS to capture additional treatments; and 2) linking oncology database records with the Mortality Information System (SIM) in order to qualify information concerning patient deaths and well as provide information for survival analyses.

The oncology database was developed and controlled by the Statistical Analysis System (SAS® 9.4). Record linkage was performed in FRIL (Fine-grained record linkage software) version 2.1.5. Descriptive statistics for treatment times for cancers considered potentially curable with early detection and treatment, as defined by Farmer et al. 2010(Farmer et al., 2010), were computed. Trend analysis of epidemiological measures for patients with these cancers were performed using Joinpoint Regression, version 4.1.1.3.

RESULTS: a Brazilian national oncology database, designed for epidemiological analysis, with information from SUS cancer patients. We obtained estimates of incidence of treatments and described trends of time between diagnosis and treatment initiation of patients with a subset of cancers potentially treatable and/or frequent.

CONCLUSION: record linkage showed to be an effective process for the creation of a national database of cancer patients in chemo- and/or radiotherapy treatment in the SUS between 2000 and 2012. APAC Oncology database concentrates important epidemiological information and can be useful to generate evidence to inform policy and services. SUS-APAC oncology database may be an important tool for the Ministry of Health, to construct indicators and monitoring, in accordance with Brazilian Strategic Action Plan to Combat Chronic Non-communicable Diseases and the global targets set to confront these diseases by 2025. Treatment of this group of cancers expanded notably over these 12 years. Time from diagnosis to treatment during this period of wider geographic reach and major expansion of coverage showed stable, emphasizing the necessity of continuous monitoring of this indicator to guide the implementation of public policies aimed at improving care of patients with cancer.

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ILUSTRAÇÕES

APRESENTAÇÃO

Este trabalho consiste na tese de doutorado intitulada “Pareamento de registros das grandes bases do SUS para permitir análises de coortes de pacientes com câncer”, apresentada ao Programa de Pós-Graduação em Epidemiologia da Universidade Federal do Rio Grande do Sul, em 22 de agosto de 2017. O trabalho é apresentado em três partes, na ordem que segue:

1. Introdução, Revisão da Literatura e Objetivos

2. Artigos

3. Conclusões e Considerações Finais.

Documentos de apoio estão apresentados nos anexos.

INTRODUÇÃO

Informações sobre as neoplasias malignas, em termos de frequência, distribuição geográfica e temporal, e de acordo com as localizações tumorais e seu estadiamento, constituem a base fundamental de um sistema de vigilância do câncer (Jacob Kligerman, 2002). O conhecimento sobre o acesso e a utilização dos serviços de atenção oncológica por parte de grupos populacionais específicos pode fornecer subsídios para as diferentes esferas de governo, possibilitando a definição de estratégias conjuntas que favoreçam o diagnóstico e tratamento, mais efetivo e com maior equidade desse agravo (Silva and Mattos, 2012). No Brasil, as bases de dados para delinear o perfil epidemiológico do câncer são oriundas do Sistema de Informação de Câncer (SISCAN), Sistema de Informações sobre Mortalidade (SIM) e Registros de Câncer de Base Populacional (RCBP) e Registros Hospitalares de Câncer (RHC)(Moura, Curado, Simões, Cezário, and Urdaneta, 2006)(Ministério da Saúde Brasil, 1996).

Contudo, o Ministério da Saúde possui bases de dados de grande magnitude com informações dos tratamentos realizados em oncologia para população usuária do Sistema Único de Saúde (SUS), incluindo todos os tratamentos cirúrgicos, quimioterápicos, radioterápicos e cuidados paliativos para neoplasia maligna ou benigna. Essas informações estão presentes nas bases compostas por dados de Autorização de Procedimentos de Alta Complexidade (APAC) e de Autorização de Internações Hospitalares (AIH)(Bittencourt, Camacho, and Leal, 2006).

Considerando a magnitude e importância do câncer para a saúde pública, o governo brasileiro regulamentou a Lei nº 12.732/12(Casa Civil Brasil, 2013) (em vigor desde 23/05/2013), a qual estabelece que o primeiro tratamento oncológico no SUS deve se iniciar no prazo máximo de 60 dias a partir da assinatura do laudo patológico ou em prazo menor conforme necessidade terapêutica do caso registrada no prontuário do paciente.

Este trabalho visa construir bases nacionais de pacientes em tratamento de câncer, utilizando os dados do Sistema de Informações Ambulatoriais de Saúde (SIA) – a saber, Autorização para Procedimentos de Alta Complexidade (APAC) e Sistema de Informações Hospitalares (SIH) - para definir o tempo de espera entre o diagnóstico e o início de tratamento de cânceres passíveis de prevenção,

detecção precoce e tratamento (Farmer et al., 2010) na população usuária do SUS. Ele permitirá também explorações futuras sobre a sobrevida relacionada aos tipos de câncer estudados, vinculando as bases obtidas com os dados do Sistema de Informação Sobre Mortalidade - SIM.

RELEVÂNCIA E JUSTIFICATIVA

As doenças crônicas não transmissíveis (DCNT) são a principal causa de mortalidade, carga de doenças e custos em saúde na grande maioria dos países no mundo (Organisation mondiale de la santé, Alwan, and Agis, 2011). Em países de média e baixa renda as taxas de mortalidade decorrentes de DCNT, ajustadas para idade, são consideravelmente mais elevadas que em países de alta renda(WHO | Global Health Observatory Data Repository [Internet], 2011) – onde se observa uma redução dessas taxas nas últimas décadas (World Health Organization, 2011). Os registros de mortalidade e DCNT disponíveis em países de média ou baixa renda apresentam problemas de consistência ou, numerosos e desunificados, se restringem a cobrir a informação de pequenas áreas, o que dificulta a exploração de tendência para taxas de mortalidade e sobrevida (Coleman et al., 2008).

Os problemas decorrentes das DCNT foram discutidos mundialmente em 2011 na Reunião de Alto Nível da Organização das Nações Unidas (ONU), tendo também como foco a realidade de países de média e baixa renda(World Health Organization (WHO), 2011). Em resposta a esse desafio colocado pelas DCNT, o Brasil desenvolveu o Plano de Ações Estratégicas para o Enfrentamento das DCNT no Brasil 2011-2022 (Ministério da Saúde, 2010). Esse plano apresenta à sociedade civil as metas e a implementação de políticas públicas para prevenção e controle das DCNT, tendo como foco as quatro principais doenças (cardiovascular, câncer, doença respiratória crônica e diabetes)(WHO | 2008-2013 Action plan for the global strategy for the prevention and control of noncommunicable diseases [Internet], 2009).

A OMS propõe, como uma das metas fundamentais, a redução em 2% ao ano na probabilidade incondicional de se ir a óbito prematuramente por um dos quatro principais grupos de DCNT(World Health Organization (WHO), 2011). No período de 2000 a 2011, o Brasil apresentou um declínio de cerca de 2,5% ao ano para esse indicador – existindo queda significativa para todas as regiões brasileiras (Malta

et al., 2014*a*) (Brasil. Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Análise de Situação de Saúde*., 2011). A taxa de mortalidade por câncer, umas das quatro principais DCNT, apresentou declínio de 0,9% - aquém das outras DCNT, porém um declínio constante ao longo dos anos desse período(Malta et al., 2014*a*).

O Ministério da Saúde está definindo ações, como a alocação de novas unidades de quimioterapia e radioterapia no País, e deseja documentar a situação de tratamento de casos atualmente diagnosticados. Na esfera legislativa, como ação para promoção da saúde pública, o governo brasileiro regulamentou a Lei nº 12.732/12(Casa Civil Brasil, 2013) (em vigor desde 23/05/2013), a qual estabelece que o primeiro tratamento oncológico no SUS deve se iniciar no prazo máximo de 60 dias a partir da assinatura do laudo patológico ou em prazo menor conforme necessidade terapêutica do caso registrada no prontuário do paciente. Posteriormente, variáveis como o tempo entre detecção e tratamento desses cânceres serão indicadores de processo para a avaliação das ações instituídas.

O pareamento, ou relacionamento, de bases de dados, apresenta crescente desenvolvimento tanto em termos de técnica(Winkler, 2006) como em popularização. Essa popularização se deve ao aumento da capacidade de processamento dos computadores(Jaro, 1995; Jin, Li, and Mehrotra, 2003; Winkler, 2006). O emprego desse procedimento para obtenção de informações na pesquisa em saúde também se mostra em franca expansão(Coeli and Camargo Jr., 2002).

As bases de dados do SUS representam um grande recurso para a vigilância de diversas doenças. O pareamento entre essas bases de dados permite descrever importantes indicadores e caracterizar correlatos populacionais, diferenças geográficas e tendências temporais, bem como os custos ao governo no cuidado dessas doenças.

Dessa forma, considerando-se o impacto das DCNT e a necessidade do cumprimento das metas e da implementação de políticas públicas por parte do governo brasileiro, o desenvolvimento de pareamento entre grandes bases do SUS para gerar informações essenciais para vigilância de doenças, afigura-se como um passo importante no enfrentamento das DCNT. Mais especificamente em relação ao câncer,

esse trabalho permitirá descrever o tempo entre a detecção e o tratamento, um indicador de processo das ações instituídas nacionalmente.

REVISÃO DE LITERATURA

1. Câncer no Brasil e no Mundo

O câncer é hoje a segunda maior causa de morbidade e mortalidade em todo o mundo, sendo responsável pela morte de 8,8 milhões de pessoas em 2015. A doença não se distribui igualmente quanto ao desenvolvimento econômico, apresentando maior mortalidade em países de renda acima da média: cerca de 70% de todas as mortes por câncer ocorrem em países de baixa e média renda, e esses respondem por 84% da população mundial(World Bank, 2016; Trading Economics - World Bank, 2017). Nos países com médio ou baixo índice sócio demográfico(GBD 2015 Mortality and Causes of Death Collaborators, 2016) os tipos de câncer mais frequentes na população masculina foram próstata, traqueia-brônquios-pulmões, estomago e cólon e reto; ao passo que para população feminina são os cânceres de mama, estômago, cólon e reto e traqueia-brônquios-pulmões (Fitzmaurice et al., 2017). Nesses países, quando comparados com o grupo de países de índice sócio demográfico acima da média, as taxas de sobrevivência são menores, devido ao diagnóstico acontecer em estádios mais avançados da doença(INCTR The International Network For Cancer Treatment and Research, 2017).

Para o ano de 2015, foram observados 17,5 milhões de casos novos de câncer em todo planeta – sendo 8,2 milhões (46,9%) entre mulheres e 9,3 milhões (53,0%) entre homens, atingindo uma razão aproximada de homem/mulher de 10:9.

A nível mundial, para ambos os sexos, os dez cânceres mais incidentes são: 1) mama (2,4 milhões de novos casos, 534 mil mortes); 2) traqueia-brônquios-pulmões (2 milhões de novos casos, 1,7 milhão de mortes); 3) cólon e reto (1,6 milhão de novos casos, 376 mil mortes); 4) próstata (1,6 milhão de novos casos, 364 mil mortes); 5) estômago (1,3 milhão de novos casos, 819 mil mortes); 6) fígado (854 mil novos casos, 810 mil mortes); 7) linfomas não-Hodgkin (666 mil novos casos, 231 mil mortes); 8) leucemia (606 mil novos casos, 353 mil mortes); 9) bexiga (540 mil novos casos, 188 mil mortes); e 10) útero (455 mil novos casos, 239 mil mortes).

Na Europa, a França tem a maior incidência de câncer entre homens – 385 por 100.000 habitantes – e a Dinamarca a maior taxa de incidência entre as mulheres – 328 por 100.000 habitantes. As taxas mundiais de incidência de câncer padronizadas por idade são de 304,6 casos por 100.000 habitantes entre homens e 229.2 casos por 100.000 habitantes entre mulheres.

O número de casos novos para todos cânceres apresenta uma tendência crescente de 33% no mundo quando avaliamos o período 2005 a 2015. O crescimento populacional é responsável por 12,6% desse crescimento, alteração na estrutura etária por 16,4% e alterações nas taxas de incidência por 4,1%. O cenário segue alarmante quando observados os dados para os cânceres mais incidentes entre homens e entre as mulheres: câncer de próstata com um aumento de 66,1% na incidência, seguido do câncer de mama com aumento de 43% na incidência entre 2005 e 2015 (Fitzmaurice et al., 2017).

Os fatores de complexidade socioeconômica atuam sobre as taxas de incidência padronizadas para idade, as quais apresentam um padrão de variação que acompanha o Índice de Desenvolvimento Humano (*Human Development Index* – IDH). Entre os homens, as taxas de incidência se comportam da seguinte maneira entre os níveis gerais do IDH: em países com IDH muito alto, 316 casos por 100.000 habitantes e 103 casos por 100.000 para países com baixo IDH; para as mulheres, 253 casos por 100.000 em países de IDH muito alto, e 123 casos por 100.000 em países com baixo IDH. Esses dados em função do IDH valem para o ano de 2012 (J. Ferlay, I. Soerjomataram, and M. Ervik, 2012).

A carga do câncer continuará aumentando nos países em desenvolvimento e crescerá ainda mais em países desenvolvidos se medidas preventivas não forem amplamente aplicadas. As projeções apresentadas por (Bray, Jemal, Grey, Ferlay, and Forman, 2012) mostram que teremos em média 23,6 milhões de casos novos de câncer a cada ano até 2030, se mantidas as tendências de incidência dos principais cânceres e mantido o atual crescimento populacional observado e projetado para o futuro. Isso significa um aumento de 68% em relação a dados observados para 2012. Tendo por base o IDH para análise, o aumento previsto é de 66% entre países com baixo ou médio IDH, e 56% entre países com IDH alto ou muito alto(Bray, Jemal, Grey, Ferlay, and Forman, 2012). Outro fator que contribui para o

aumento da carga de câncer é a tendência positiva do envelhecimento da população, projeções da continuidade desse padrão, sendo o maior determinante desse aumento (Fitzmaurice et al., 2017).

No Brasil, a estimativa para os anos de 2016 e 2017, dada pelo INCA, aponta para a ocorrência de aproximadamente 600 mil casos novos de câncer, incluindo os casos de pele não melanoma. Os tipos mais incidentes de câncer, para o ano de 2016 e aplicáveis também ao ano de 2017, entre os homens, serão os cânceres de próstata (28,6%), pulmão (8,1%), intestino (7,8%), estômago (6,0%) e cavidade oral (5,2%). Já entre as mulheres, os cânceres mais incidentes serão o de mama (28,1%), intestino (8,6%), colo do útero (7,9%), pulmão (5,3%) e estômago (3,7%). Estimam-se 395 mil casos novos de câncer – 204 mil para o sexo masculino e 190 mil para sexo feminino – para o ano de 2014, isso sem considerar os casos de câncer de pele não melanoma. O câncer de pele do tipo não melanoma será o mais incidente na população brasileira com 176 mil, sendo 80.850 casos novos de câncer de pele não melanoma nos homens e 94.910 nas mulheres no Brasil, em 2016 (Instituto Nacional de Câncer, 2016).

2. As DCNT como foco de Saúde Pública

As doenças crônicas não transmissíveis (DCNT) são a principal causa de mortalidade, carga de doenças e custos em saúde na grande maioria dos países no mundo (Organisation mondiale de la santé, Alwan, and Agis, 2011). Em países de média e baixa renda as taxas de mortalidade decorrentes de DCNT, ajustadas para idade, são consideravelmente mais elevadas que em países de alta renda (WHO | Global Health Observatory Data Repository [Internet], 2011) – onde se observa uma redução dessas taxas nas últimas décadas (World Health Organization, 2011). Os registros de mortalidade e DCNT disponíveis em países de média ou baixa renda apresentam problemas de consistência ou, numerosos e desunificados, se restringem a cobrir a informação de pequenas áreas, o que dificulta a exploração de tendência para taxas de mortalidade e sobrevida (Coleman et al., 2008). Do ponto de vista econômico, as DCNT impactam como grande custo aos sistemas de saúde dos países, a sua sociedade, atingindo o nível familiar. A OMS aponta que essas doenças criam um círculo vicioso com a pobreza, impactando negativamente sobre o desenvolvimento

macroeconômico dos países, fundamentalmente daqueles de média e baixa renda (Abegunde, Mathers, Adam, Ortegon, and Strong, 2007; Malta et al., 2014*b*).

Em 2008, a Organização Mundial da Saúde (OMS) estimou que 63% das mortes globais se devem às DCNT (Malta, Jr, and Da, 2013), totalizando aproximadamente 36 milhões de mortes. Destacaram-se como principais causas de morte o aparelho circulatório, diabetes, câncer e doença respiratória crônica (Malta et al., 2014*b*).

No Brasil, as DCNT foram responsáveis por 75.8% da mortalidade total no país em 2015, atingindo uma taxa de mortalidade ajustada para idade de 611 óbitos por 100 mil habitantes (GBD 2015 Mortality and Causes of Death Collaborators, 2016; Malta et al., 2017). Com base em dados do SIM, em 2015, corrigidos erros de sub-registro e realizada redistribuição das causas mal definidas de óbito, ocorreram 1.029 milhão de óbitos atribuídos às DCNT. A taxa de mortalidade ajustada para idade e agregada para todas DCNT, e para toda população do país, mostrou uma redução de 25.3% entre os anos de 1990 e 2015. Entre as quatro principais DCNT, centro de atenção dos planos de combate às DCNT, para o mesmo ano, as doenças cardiovasculares foram responsáveis por 424.058 mortes (41.2%), as neoplasias por 236.345 mortes (23%), as doenças respiratórias por 79.651 (7,7%) e o diabetes mellitus por 62.466 mortes (6,1%) entre todas DCNT. No que tange tendências da taxa de mortalidade padronizada para idade, o quadro para DCNT no Brasil é promissor, uma vez que, na comparação entre os anos de 1990 e 2015, constatou-se uma redução de 40,5% para doença cardiovasculares, redução de 28,9% para doenças respiratórias e estabilidade para as neoplasias e o diabetes mellitus (Malta et al., 2017).

A Organização das Nações Unidas (ONU), em setembro de 2011, em função da carga das DCNT e seu impacto sobre os sistemas de saúde e a sociedade, convocou uma reunião de alto nível sobre DCNT. Nessa reunião foram debatidos compromissos a serem assumidos pelos países-membros para deter o crescimento das DCNT, estabelecendo ações de prevenção de seus principais fatores de risco e empenho por uma melhor atenção à saúde (Malta et al., 2014*b*). Em 2013, a OMS lançou o Plano Global de Ação para a Prevenção e Controle de Doenças Não Transmissíveis para 2013-2020 (*Global Action Plan for the Prevention and Control of*

Noncommunicable Diseases 2013-2020), que tem por objetivo a redução de 25% das mortes prematuras por câncer, doenças cardiovasculares, diabetes e doenças respiratórias crônicas até 2025. Entre as ações contra o câncer, o plano prevê, pela meta número 5, a redução da prevalência do consumo de tabaco em 30% (Bray, Jemal, Grey, Ferlay, and Forman, 2012).

No sentido de implementar as definições do plano de ação citado acima, a Organização Mundial de Saúde (OMS), a Agência Internacional para Pesquisa sobre Câncer (*International Agency for Research on Cancer* – IARC) em colaboração com a Organização das Nações Unidas (ONU) estão realizando a *United Nations Noncommunicable Diseases Interagency Taskforce* (2014) para a ampliar o comprometimento político entre países para a prevenção e controle do câncer. A força tarefa ainda atua no sentido de: coordenar e realizar pesquisas sobre as causas do câncer em humanos e os mecanismos da carcinogênese; monitorar a carga de câncer como parte do trabalho da Iniciativa Global sobre Registros de Câncer - *Global Initiative on Cancer Registries*, GICR); desenvolver estratégias científicas para a prevenção e controle do câncer; gerar novos conhecimentos, e disseminar o conhecimento existente para facilitar a entrega de abordagens baseadas em evidências para o controle do câncer; desenvolver padrões e ferramentas para orientar o planejamento e implementação de intervenções para a prevenção, detecção precoce, tratamento e cuidados; facilitar grandes redes de parceiros de controle do câncer e peritos a nível mundial, regional e nacional; fortalecer os sistemas de saúde a nível nacional e local para entregar tratamentos e cuidados para pacientes com câncer; e prestar assistência técnica para uma rápida, efetiva transferência das intervenções de melhores práticas para os países em desenvolvimento (Organização Mundial de Saúde, 2014).

3. Compromissos assumidos pelo governo brasileiro para DCNT

O Brasil participou ativamente dessa mobilização global ao lançar o Plano de Ações Estratégicas para o Enfrentamento das Doenças Crônicas Não Transmissíveis (DCNT) no Brasil, 2011-2022 (Ministério da Saúde, 2010), que define metas e compromissos, ações e investimentos no sentido de preparar o país para o enfrentamento dos desafios representados pelas DCNT e seus fatores de risco nos

próximos dez anos. O plano tem como foco as quatro principais doenças (cardiovascular, câncer, doença respiratória crônica e diabetes) e seus fatores de risco (tabagismo, consumo nocivo de álcool, inatividade física, alimentação inadequada) (WHO | 2008-2013 Action plan for the global strategy for the prevention and control of noncommunicable diseases [Internet], 2009). Destaca-se, entre as metas traçadas, a redução de 2% ao ano na mortalidade prematura devida às quatro principais causas de mortalidade por DCNT, destacadas pela OMS e focalizadas no Plano (Malta et al., 2014*b*). Numa visão heurística, o plano apresenta três diretrizes fundamentais: 1) vigilância, informação, avaliação e monitoramento; 2) promoção da saúde; e 3) cuidado integral.

No que diz respeito aos temas de morbimortalidade, fatores de risco de DCNT e sistemas de saúde, o Plano de Ações Estratégicas para o Enfrentamento das DCNT considera as seguintes metas nacionais: 1) redução da taxa de mortalidade prematura (<70 anos) por DCNT em 2% ao ano; 2) redução da prevalência de excesso de peso e obesidade em crianças de 5 a 9 anos; 3) redução da prevalência de excesso de peso e obesidade em adolescentes de 10 a 19 anos; 4) deter o crescimento de excesso de peso e obesidade em adultos (≥18 anos); 5) redução das prevalências de consumo nocivo de álcool, de 18% (2011) para 12% em (2022); 6) redução da prevalência de tabagismo em adultos, de 15,1% (2011) para 9,1% em (2022); 7) aumento da prevalência de atividade física no lazer, de 14,9% (2010) para 22% (2022); 8) aumento do consumo de frutas e hortaliças, de 18,2% (2010) para 24,3% (2022); 9) redução do consumo médio de sal, de 12 gramas (2010) para 5 gramas (2022); 10) aumento da cobertura de mamografia em mulheres entre 50 e 69 anos, de 54% (2008) para 70% (2022); 11) ampliação da cobertura de exame preventivo de câncer de colo uterino em mulheres de 25 a 64 anos, de 78% (2008) para 85% (2022); 12) garantia do tratamento de mulheres com diagnóstico de lesões precursoras de câncer de colo de útero e de mama(Malta, Jr, and Da, 2013).

Para vigilância, os indicadores de monitoramento são: morbidade e mortalidade (monitoramento da incidência do câncer); fatores de risco (redução do consumo de ácidos graxos saturados e aumento do consumo de cinco porções de frutas e hortaliças por dia); e respostas dos sistemas nacionais. A vigilância das respostas dos sistemas nacionais englobam cuidados paliativos de câncer; políticas nacionais de

regulamentação de alimentação saudável; disponibilidade de vacinas contra o papiloma vírus humano (HPV) mediante comprovação de custo-eficácia e de acordo com programas e políticas nacionais, bem como políticas de regulamentação da propaganda infantil de alimentos, rastreamento de câncer cervical e vacinação contra o vírus da hepatite B (Ministério da Saúde, 2010).

A OMS propõe, como uma das metas fundamentais, a redução em 2% ao ano na probabilidade incondicional de se ir a óbito prematuramente por um dos quatro principais grupos de DCNT(World Health Organization (WHO), 2011). No período de 2000 a 2011, o Brasil apresentou um declínio de cerca de 2,5% ao para esse indicador – existindo queda significativa para todas as regiões brasileiras (Brasil. Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Análise de Situação de Saúde*., 2011; Malta et al., 2014*a*) A taxa de mortalidade por câncer, umas das quatro principais DCNT, apresentou declínio de 0,9% - aquém das outras DCNT, porém um declínio constante ao longo dos anos desse período (Malta et al., 2014*a*).

4. Câncer como uma doença prioritária

O câncer é hoje também, no Brasil, um problema de saúde pública, cujos controle e prevenção deverão ser priorizados em todas as regiões, desde as mais desenvolvidas – cultural, social e economicamente – até às mais desiguais. As abordagens orientadas para enfrentar esse problema de saúde são, necessariamente, múltiplas, incluindo: ações de educação para saúde em todos os níveis da sociedade; prevenção orientada para indivíduos e grupos; geração de opinião pública; apoio e estímulo à formulação de legislação específica para o enfrentamento de fatores de risco relacionados à doença; e fortalecimento de ações em escolas e ambientes de trabalho (Instituto Nacional de Câncer José Alencar Gomes da Silva and Coordenação de Prevenção e Vigilância, 2014).

Para termos uma ideia do impacto econômico do câncer nas contas do governo – ou seja, no bolso dos pagadores de impostos – segundo o relatório do Tribunal de Contas da União, os gastos federais com tratamentos ambulatoriais e hospitalares de câncer têm apresentado uma tendência de aumento nos últimos anos – o total dos tratamentos ambulatoriais e hospitalares atingiu o montante de R$ 1,48

bilhão em 2008, passou a R$ 1,69 bilhões em 2009 e superou R$ 1,92 bilhões em 2010 (TRIBUNAL DE CONTAS DA UNIÃO, 2011). O investimento no SUS para oncologia, em 2012, seguindo trajetória crescente, foi de R$ 2,4 bilhões. Este crescimento de 26%, entre 2010 e 2012, resultaram na ampliação de 17,3% no número de sessões de radioterapia, aumentando de 7,6 milhões para 9,2 milhões. Na mesma onda de crescimento, os tratamentos de quimioterapia passaram de 2,2 milhões para 2,5 milhões - 14,8% acréscimo (Notícias, 2014). Em 2012, pelo SUS, foram realizadas mais de 500 mil internações para tratamento do câncer, ao custo de R$ 806 milhões (INCA, 2013). O dado mais recente sobre os gastos com o tratamento da doença fornecido pelo governo federal, carecendo de disponibilidade do documento ao público, é de R$3,5 bilhões em 2015. Os cálculos comportam recursos para procedimentos como cirurgias oncológicas, quimioterapia, radioterapia, hormonioterapia e cuidados paliativos. Esses dados ainda mostram a tendência crescente no número de pessoas atendidas na rede pública neste período, passando de 292 mil para 393 mil (Estadão, 2016*a*).

4.1 Tendências de mortalidade por câncer no Brasil e no Mundo

Nas últimas décadas, a mortalidade relacionada ao câncer apresenta tendência de declínio em países como Estados Unidos, Reino Unido, França e Austrália. A queda na incidência observada nos EUA, dos anos 2000 até 2010, é atribuída à redução da terapia de reposição hormonal e a diminuição do número de casos pré-clínicos detectados através de triagem, num esforço iniciado há mais de 20 anos (Jemal, Center, DeSantis, and Ward, 2010; Girianelli et al., 2014). Entretanto, em países menos desenvolvidos, ou países em transição econômica, observou-se a ascensão da taxa de mortalidade padronizada por idade para cânceres como câncer de pulmão, colo retal, feminino de mama e próstata.

Entre os anos de 2005 e 2015, observou-se um decréscimo das taxas de mortalidade padronizadas para idade para os cânceres em geral. Esse fenômeno ocorreu em 140 de 195 países no mundo, sendo o continente africano a principal região onde essa redução não foi observada. Entre os cânceres com redução no número de mortes entre 2005 e 2015 estão os canceres de esôfago, estomago, leucemia mieloide crônica e linfomas não-Hodgkin. Entre os fatores desse fenômeno se encontra a adoção

de um padrão de comportamento não saudável como fumo, falta de atividade física e consumo de alimentos altamente calóricos (Fitzmaurice et al., 2017).

No Brasil, a análise das tendências temporais da taxa de mortalidade padronizada por idade, entre 1980 e 2006, revelam uma queda do câncer cervical e um aumento para o câncer de mama a nível nacional, mas esse padrão não se verifica quando são buscadas as tendências a nível de capitais e demais municípios desagregados. Essas tendências não mostraram ter diferentes inclinações nessas desagregações. Para o plano de regiões brasileiras, entre 1980 e 2006, observou-se o aumento da taxa de mortalidade padronizada por idade por câncer de mama em todas as cinco regiões do Brasil, embora seja significativa uma tendência de queda nas taxas em capitais a partir do final da década de 1990 em diante (Girianelli et al., 2014). Já no período de 1996 a 2010, todos cânceres em conjunto, entre mulheres apresentam uma estabilidade, com uma tendência de aumento percentual de 0,4%. Por outro lado, entre homens, a tendência de aumento percentual é de 1,2% até 2008 e então seguida por uma redução de 2,1%(Barbosa, de Souza, Bernal, and Costa, 2015).

Em 2015, o Brasil registrou 205.998 mortes relacionadas ao câncer (CID-10 C00-C99), sendo 109.186 entre homens e 96.721 entre as mulheres. O câncer de brônquios e pulmão foi o tipo que mais matou (26.400 pessoas), seguido do câncer de mama (15.593), próstata (14.484), estômago (14.265), e cólon (14.265) para ambos os sexos (DATASUS, 2017).

Os dados do GBD, em comparação dos anos 1990 e 2015, mostram que a taxa de mortalidade padronizada de câncer para população de homens e mulheres permaneceu estável – tanto a nível nacional quanto a nível de unidades da federação. Aumentos da taxa de mortalidade padronizada foram observados na região nordeste e metade da região norte, porém não foram estatisticamente significativos. Neste período, observou-se redução das taxas de mortalidade padronizadas para câncer de estômago (-38,9% para mulheres, -37,3% para homens), colo do útero (-33.9%), câncer de pulmão (-12%) e esôfago (-14.1%) entre homens. Observou-se estabilidade das taxas de mortalidade de câncer de próstata ainda para o mesmo período, uma estabilidade válida tanto a nível nacional quando a nível estadual. Aumentos nas taxas de mortalidade foram observados para câncer de cólon-reto entre homens (+29.5%) e, entre mulheres, o câncer de pulmão (+20.7%).

Uma das limitações dada pela análise com um intervalo de 15 anos (1990-2015), baseada em apenas 2 pontos é não permitir uma análise da evolução da tendência da taxa de mortalidade padronizada (GBD 2015 Mortality and Causes of Death Collaborators, 2016; Guerra et al., 2017). De 2000 a 2011, observou-se um declínio de 0,9% ao ano da mortalidade relacionada ao câncer – de 147,9 por 100.000 habitantes para 132,9 por 100.000 habitantes) (Malta et al., 2014*b*).

4.2 O tratamento de câncer

Em 2010, Farmer et. al. publicou um trabalho (Farmer et al., 2010) no sentido dar força a um movimento para a expansão da atenção ao tratamento e controle do câncer em países em desenvolvimento. Eles procuram chamar atenção para as substanciais inequidades existentes entre países desenvolvidos e em desenvolvimento quanto às taxas de sobrevivência para o câncer.

Afirmam também a necessidade de estabelecer um conjunto de estratégias para a prevenção de novos cânceres e a redução dos fatores de risco, para que se reduza a disparidade entre os países em relação à sobrevida do câncer e os efeitos da doença sobre sofrimento humano. Em países com recursos limitados, sem serviços especializados, algumas experiências mostram que muito pode ser feito para prevenir e tratar o câncer, como implantação dos cuidadores primários e secundários, uso de medicamentos sem patente e aplicação de mecanismos regionais e globais para financiamento e de aquisição de medicamentos. Estas estratégias podem reduzir custos, aumentar o acesso aos serviços de saúde e reforçar os sistemas de saúde para enfrentar o desafio de câncer e outras doenças. A *Global Task Force on Expanded Access to Cancer Care and Control in Developing Countries* (GTF.CCC), formada em 2009, tem a finalidade de se dedicar para controle e expansão do acesso ao tratamento de câncer, pela implementação e avaliação de estratégias, em países em desenvolvimento. Essa força tarefa é composta por líderes das comunidades de saúde e assistência global do câncer. As propostas do GTF.CCC para países de média e baixa renda, são pautadas em experiências de sucesso avaliado como tratamentos implementados em países como Malawi, Ruanda e Haiti; a inclusão do tratamento de câncer em programas de seguro de saúde nacional do México e Colômbia; e a expansão

do acesso ao tratamento com a construção de centro de excelência em câncer, como o *King Hussein Cancer Center* na Jordânia (Farmer et al., 2010).

4.3 A política de saúde para o câncer proposta por GTF.CCC

A grande proposta do GTF.CCC é de que o controle e tratamento do câncer se tornem disponíveis nos países de média e baixa renda no menor tempo, e com a maior cobertura possíveis. Para isso, o foco do projeto para controle e tratamento do câncer desses países é fundamentalmente sobre cânceres que podem ser prevenidos ou curados, ou possivelmente atenuados. Três mudanças iniciais formam o conjunto de ação proposto. A primeira mudança é a execução dos programas para treinamento de profissionais de saúde e profissionais associados, uso de tecnologia e telecomunicações para superar limitações locais dos recursos. Isso simultaneamente associado ao monitoramento e avaliação da atenção à saúde em relação ao câncer, permitindo assim a identificação de medidas mais efetivas para reduzir impactos do câncer e ampliar os serviços de saúde nos países em desenvolvimento – o que amplia o conhecimento a ser compartilhado com outros sistemas de saúde. A segunda mudança é a definição e implementação de mecanismos de preços (regional e global) para contratos de compra coletivos de medicamentos e vacinas, bem como contratação de serviços – ou seja, formação de grupos de países para negociação em blocos para garantir aquisições com preços reduzidos. A terceira mudança, a identificação e implementação de mecanismos de financiamento inovadores, com a finalidade de expandir os recursos financeiros disponíveis para a prevenção, tratamento e cuidados paliativos de câncer nesses países em desenvolvimento (Farmer et al., 2010).

4.4 Tipos de cânceres passíveis de prevenção, detecção precoce e tratamento definidos por GTF.CCC

O GTF.CCC, reforçado por Farmer e colaboradores, apresentam a seguinte lista de cânceres passíveis de prevenção, detecção precoce e tratamento em países de baixa e média renda, sobre os quais deve existir intervenção concentrada nas ações nacionais contra o câncer.

1) Cânceres passíveis de prevenção por fator de risco: câncer de pulmão, câncer de cabeça e pescoço, câncer de bexiga associados ao tabaco; câncer do

colo do útero, câncer de cabeça e pescoço – relacionado a infecção pelo papilomavírus humano (vírus HPV); câncer hepatocelular ligado à infecção por hepatite;

2) Cânceres que são potencialmente curáveis com detecção precoce e tratamento, incluindo cirurgia: câncer do colo do útero, câncer de mama, câncer do colon-reto;

3) Cânceres que são potencialmente curáveis com tratamento sistemático e, para os quais, a detecção precoce não é crucial: linfoma de Burkitt, linfoma de grandes células (*Large-cell lymphoma*), linfoma de Hodgkin, câncer de testículo, leucemia linfoblástica aguda, sarcoma de tecido mole (*soft tissue sarcoma*), osteossarcoma;

4) Cânceres que podem ser atenuados com tratamento sistêmico: sarcoma de Kaposi, câncer de mama avançado, câncer de ovário, leucemia mielóide crônica (Farmer et al., 2010).

5. O SUS e seus compromissos constitucionais

O direito à saúde é definido na Constituição Federal de 1988 e está inserido no conjunto dos direitos sociais constitucionalmente garantidos. Trata-se de um direito público subjetivo, uma prerrogativa jurídica indisponível assegurada à totalidade da população brasileira (Lenir Santos, 2005; Elisangela Santos de Moura, 2013). Assim se encontra escrito na constituição:

"Art. 196. A saúde é direito de todos e dever do Estado, garantido mediante políticas sociais e econômicas que visem à redução do risco de doença e de outros agravos e ao acesso universal e igualitário às ações e serviços para sua promoção, proteção e recuperação" (Brazil and Brazil Congresso Nacional Câmara dos Deputados, 2003).

Esse preceito é complementado pela lei 8.080/90, em seu artigo 2º, assim escrita: "A saúde é um direito fundamental do ser humano, devendo o Estado prover as condições indispensáveis ao seu pleno exercício"(Constituição Brasil, 1990).

O artigo 196 da constituição brasileira ou a lei 8.080/90 (Constituição Brasil, 1990) não apresentam o que seja o conceito de saúde. A definição de saúde é dada pelo jurista Henrique Hoffmann Monteiro Castro (Henrique Hoffmann Monteiro de Castro, 2005): "Corresponde a um conjunto de preceitos

higiênicos referentes aos cuidados em relação às funções orgânicas e à prevenção das doenças. Em outras palavras, saúde significa estado normal e funcionamento correto de todos os órgãos do corpo humano, sendo os medicamentos os responsáveis pelo restabelecimento das funções de um organismo eventualmente debilitado".

O estado brasileiro, com a finalidade de cumprir seu dever na prestação do serviço público de atendimento à saúde, balizado pela Lei nº 8.080/90, somados aos artigos 196 a 200 da Constituição Federal de 1988, regulamentam o Sistema Único de Saúde (SUS). O SUS foi concebido como um sistema, um conjunto cujas partes encontram-se coordenadas entre si, funcionando como uma estrutura organizada, submetida a princípios e diretrizes fixados legalmente. Assim ele forma uma rede regionalizada e hierarquizada de ações e serviços de saúde (Lenir Santos, 2005).

O SUS deve atuar em razão do disposto no art. 200 da Constituição Federal, que afirma a saúde como direito de todos brasileiros e dever do Estado. Assim se encontra escrito na constituição:

"Art. 200. Ao sistema único de saúde compete, além de outras atribuições, nos termos da lei:

I - controlar e fiscalizar procedimentos, produtos e substâncias de interesse para a saúde e participar da produção de medicamentos, equipamentos, imunobiológicos, hemoderivados e outros insumos;

II - executar as ações de vigilância sanitária e epidemiológica, bem como as de saúde do trabalhador;

III - ordenar a formação de recursos humanos na área de saúde;

IV - participar da formulação da política e da execução das ações de saneamento básico;

V - incrementar em sua área de atuação o desenvolvimento científico e tecnológico;

VI - fiscalizar e inspecionar alimentos, compreendido o controle de seu teor nutricional, bem como bebidas e águas para consumo humano;

VII - participar do controle e fiscalização da produção, transporte, guarda e utilização de substâncias e produtos psicoativos, tóxicos e radioativos;

VIII - colaborar na proteção do meio ambiente, nele compreendido o do trabalho" (Brazil and Brazil Congresso Nacional Câmara dos Deputados, 2003).

A Lei n. 8.080/90, em seus artigos 5º e 6º, estabelece os objetivos e as atribuições do SUS. A lei funciona explicitando o art. 200 da Constituição Federal, tendo inclusive repetição dos incisos daquele artigo. Os objetivos do SUS são os seguintes:

1) a identificação e divulgação dos fatores condicionantes e determinantes da saúde;

2) a formulação de políticas de saúde destinadas a promover, nos campos econômico e social, a redução de riscos de doenças e outros agravos; e

3) execução de ações de promoção, proteção e recuperação da saúde, integrando as ações assistenciais com as preventivas, de modo a garantir às pessoas a assistência integral à sua saúde (Lenir Santos, 2005).

A direção do SUS é única, de acordo com o inciso I do art. 198 da Constituição Federal. Em cada uma das três esferas de governo definida pela constituição de 1988, a direção do SUS é exercida pelos seguintes órgãos:

1) no âmbito da União, pelo Ministério da Saúde,
2) no âmbito dos Estados e do Distrito Federal, pelas respectivas Secretarias de Saúde ou órgãos equivalentes e
3) no âmbito dos Municípios, pelas respectivas Secretarias de Saúde ou órgãos equivalentes.

A Portaria nº 3.916 aprova a Política Nacional de Medicamentos, e a mais recente Norma Operacional da Assistência à Saúde, nº 01/2002 – NOAS-SUS 01/02, aprovada por Portaria do Ministério da Saúde e vem a suceder a Norma Operacional Básica do SUS, nº 01/96 – NOB-SUS 01/96(Lenir Santos, 2005).

Em conclusão, o Estado tem o dever de assegurar efetivamente o direito à saúde a todos os cidadãos, zelando pelo seu direito à vida. A Constituição Federal, em seus dispostos, garante o acesso universal e igualitário às ações e serviços para a promoção, proteção e recuperação da saúde, assegurando, portanto, a sua proteção nas órbitas genérica e individual.

6. Ações do governo brasileiro contra o câncer

A questão do câncer ganhou importância, no Brasil, pelas características epidemiológicas de morbidade e mortalidade, que essa doença vem apresentando, e, com isso, o tema conquista espaço nas agendas políticas e técnicas de todas as esferas de governo – como demonstra o Plano de Ações Estratégicas para o Enfrentamento das DCNT no Brasil, 2011-2022 (Ministério da Saúde, 2010). O conhecimento da situação dessa doença entre na população brasileira é fundamental para estabelecer prioridades e alocar recursos de forma direcionada para o controle desse cenário. Os 16 Objetivos Estratégicos do Ministério da Saúde para o período 2011 – 2015, no sentido de promover ações de controle de câncer, de redução da prevalência do tabagismo e de ampliação de acesso, diagnóstico e tratamento em tempo oportuno dos cânceres de mama e do colo do útero, assim como a publicação da nova Política Nacional de Prevenção e Controle de Câncer na Rede de Atenção às Pessoas com Doenças Crônicas (PNPCC-RAS), por meio da Portaria no 874, de 16 de maio de 2013 (Instituto Nacional de Câncer José Alencar Gomes da Silva and Coordenação de Prevenção e Vigilância, 2014)(Ministério da Saúde Brasil, 2013*a*), são exemplos de como o tema conquista espaço nas agendas políticas e técnicas de todas as esferas de governo (Ministério da Saúde Brasil, 2013*a*).

6.1 Esforços para prevenção no Brasil

O Global Burden of Disease Study 2015 aponta os seguintes principais fatores de risco associados à mortalidade por câncer no Brasil: tabaco (42.3 mil mortes, 3.1% de todas as mortes em geral); riscos associados à padrões alimentares (24,7 mil mortes, 1,8% de todas as mortes em geral); alto índice de massa corporal (12,1 mil mortes, 0,9% de todas as mortes em geral); uso de álcool e drogas (12 mil mortes, 0,8% de todas as mortes em geral); riscos ocupacionais (11,3 mil mortes, 0,8% de todas as mortes em geral); sexo inseguro (10 mil mortes, 0,7% de todas as mortes em geral); baixa atividade física (6,3 mil mortes, 0,4% de todas as mortes em geral); e poluição do ar e outros fatores ambientais (5,2 mil mortes, 0,4% de todas mortes em geral). Esses fatores são responsáveis por 95,5 mil mortes entre os cinco principais tipos de neoplasias malignas (a saber, canceres de traqueia, brônquios e pulmão, estômago, cólon-reto, esôfago e colo uterino) listadas entre as 20 principais causas de

morte devido a DCNT no Brasil em 2015 – 7% de todas as mortes, 45,9 mortes por 100.000 habitantes (GBD 2015 Mortality and Causes of Death Collaborators, 2016; IHME, 2016).

6.1.1 Sexo inseguro

A infecção pelo HPV é um fator necessário, mas não suficiente, para o desenvolvimento do câncer de colo do útero. Aproximadamente 291 milhões de mulheres no mundo são portadoras do HPV, sendo que 32% estão infectadas pelos tipos 16, 18 ou ambos. Comparando-se esse dado com a incidência anual de aproximadamente 500 mil casos de câncer de colo do útero, conclui-se que o câncer é um desfecho raro, mesmo na presença da infecção pelo HPV.

O Ministério da Saúde avalia desde 2006 a incorporação da vacina contra o HPV pelo SUS. Em dezembro de 2011, um resultado do estudo de custo-efetividade da incorporação da vacina contra HPV no Programa Nacional de Imunização foi realizado pelo Ministério da Saúde. Esse estudo concluiu que a vacinação seria custo-efetiva no país. A primeira reunião do grupo de trabalho para elaboração das diretrizes para introdução da vacina no calendário nacional aconteceram em julho de 2012. Em 2013, o Ministério da Saúde considerou adequado incluir o imunizante no calendário nacional e desde 2014 o SUS disponibiliza a vacina contra o HPV. É interessante ressaltar que três pareceres anteriores – de 2007, 2010 e 2011 – contraindicaram o uso da vacina contra o HPV como política de saúde. Decidiu-se então, como medida cautelosa, esperar o resultado de estudo sobre custo-efetividade da vacinação no cenário brasileiro, a avaliação do impacto na sustentabilidade do Programa Nacional de Imunizações (PNI) e as negociações para transferência de tecnologia para produção da vacina no país para subsidiar a tomada de decisão do Ministério da Saúde sobre o tema.

6.1.2 Tabagismo

O Plano de Ações Estratégicas para o Enfrentamento das DCNT no Brasil, 2011-2022 (Ministério da Saúde, 2010), tem entre seus objetivos combater o tabagismo no País. Responsável por cerca de 200 mil mortes por ano no Brasil (Instituto Nacional de Câncer José Alencar Gomes da Silva and Coordenação de Prevenção e Vigilância, 2014), o tabagismo é reconhecido, pela Organização Mundial

da Saúde (OMS), como uma doença epidêmica, e é considerado fator de risco para aproximadamente 50 doenças, principalmente as respiratórias e cardiovasculares, além de vários tipos de câncer como o de pulmão e brônquios, um câncer agressivo e que geralmente apresenta os primeiros sintomas já em estágio avançado (Stewart, Wild, International Agency for Research on Cancer, and World Health Organization, 2014).

O governo, em suas esferas federal e estatuais, aprovou a Lei Antifumo. Assim os estabelecimentos comerciais destinados especificamente à comercialização de produtos fumígenos e os ambientes fechados onde o fumo será permitido – tabacarias, locais de pesquisas e sets de filmagens – tiveram que se adequar para atender as determinações da lei. A Lei Antifumo tem como objetivo garantir a proteção à saúde dos trabalhadores expostos ao fumo. Determina então que locais comerciais e de trabalho devem ter uma área exclusiva para o consumo, com sistema de ventilação por exaustão capaz de reduzir o acúmulo de emissões de fumaça no seu interior e evitar a contaminação dos demais ambientes. Nesses ambientes, não será permitida a venda e fornecimento de alimentos e bebidas. Os fumantes, no entanto, poderão levar para o interior do local o que forem consumir.

O sistema de ventilação deverá ser mantido em operação após a desocupação e desativação da área exclusiva, sendo desligado automaticamente, para exaurir os resíduos e odores que podem permanecer no ambiente fechado. Os revestimentos, pisos, tetos e bancadas dessas áreas deverão ser resistentes ao uso de desinfetantes, com o menor número possível de ranhuras ou frestas. O mobiliário deve ser de material não combustível, de fácil limpeza e que minimize a absorção das partículas. Os serviços de limpeza e de manutenção das instalações e equipamentos só poderão ser feitos quando os locais não estiverem em funcionamento.

A regulamentação das regras para proteger o trabalhador já estava prevista no Decreto 8.262/14, popularmente conhecido como Lei Antifumo. A legislação proíbe o consumo de cigarros, cigarrilhas, charutos, cachimbos e outros produtos fumígenos, derivados ou não do tabaco, em locais de uso coletivo, públicos ou privados, mesmo que o ambiente esteja só parcialmente fechado por uma parede, divisória, teto ou até toldo. Os narguilés também estão vetados.

A norma também extingue os fumódromos e acaba com a possibilidade de propaganda comercial de cigarros nos displays dos pontos de venda.

Outra obrigatoriedade prevista é o aumento dos espaços para os avisos sobre os danos causados pelo tabaco e a presença de advertências em 30% da parte frontal das embalagens dos produtos a partir de 2016.

O Ministério da Saúde atua também contra o tabagismo com peças de sua campanha publicitária para conscientizar a população, sindicatos e proprietários de estabelecimentos comerciais sobre o início da vigência da proibição de fumar em recintos coletivos de todo país – a campanha é voltada para o público jovem (até 25 anos) e adultos. Essa campanha, além de veiculada na internet, conta com cartazes e folders para a população geral e determina que estabelecimentos que vendem produtos do tabaco apresentem a publicidade, além de apresentar as alterações que acontecerão e como elas são positivas para todos (fumantes e não fumantes)(Presidência da República, Casa Civil, and Subchefia para Assuntos Jurídicos, 2011; Ministério da Saúde Brasil, 2013*b*; Presidência da República, Casa Civil, and Subchefia para Assuntos Jurídicos, 2014; Ministério da Saúde, 2014).

No Brasil, o número de fumantes permanece em queda. Segundo o Vigitel 2016 (Vigilância de Fatores de Risco e Proteção para Doenças Crônicas por Inquérito Telefônico), o percentual caiu 28% nas capitais brasileiras nos últimos oito anos. Em 2006, 15,7% da população adulta que vive nas capitais fumava, ao passo que, em 2016, a prevalência caiu para 10,2% - entre homens 12,7%, e 8,0% entre mulheres (MINISTÉRIO DA SAÚDE BRASIL, Secretaria de Vigilância em Saúde, and Departamento de Vigilância de Doenças e Agravos não Transmissíveis e Promoção da Saúde, 2014, 2017). Essa prevalência é cerca de três vezes menor que o índice de 1989, quando a Pesquisa Nacional de Saúde e Nutrição (PNSN), realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE), mostrou que 34,8% de fumantes na população (Instituto Nacional de Alimentação e Nutrição, 1989). A meta do Ministério da Saúde é chegar a 9% nas capitais até 2022.

No Brasil, o tratamento do tabagismo está vinculado ao SUS e é regulado pela Portaria do Ministério da Saúde nº 571 publicada em 05 de abril de 2013 (Portaria nº 571/2013)(Ministério da Saúde Brasil, 2013*c*). O SUS oferece o tratamento para quem deseja parar de fumar contando com, 23.387 equipes da família, em 4.375 municípios, que estão preparadas para atender à essa população. Além do acompanhamento profissional, são oferecidos medicamentos, como adesivos,

pastilhas, gomas de mascar e o bupropiona. O Ministério da Saúde destinou R$ 41 milhões para compra desses medicamentos, o que permitiu o tratamento de mais de 145 mil tabagistas em 2014 (Instituto Nacional de Câncer José Alencar Gomes da Silva, 2014*a*) (Instituto Nacional de Câncer José Alencar Gomes da Silva, 2014*b*).

6.2 Esforços para ampliar o acesso ao tratamento

O governo brasileiro, para solucionar o problema da a falta de equipamentos para radioterapia, fundamentalmente nas regiões Norte e Nordeste, e reduzir a existência de uma não homogeneidade na distribuição geográfica das unidades de serviço – chamadas "vazios no atendimento" – em todas as demais regiões, apontadas por relatório do Tribunal de Contas da União, realizou a aquisição de novos equipamentos para radioterapia. Em novembro de 2013, o Ministério da Saúde anunciou a aquisição de 80 aceleradores lineares, para a realização de radioterapia, para serem distribuídos em 22 estados e no Distrito Federal – dessa forma ampliando em 25% da oferta desse tipo de tratamento no SUS quando totalmente em funcionamento. Segundo definição em conjunto do Ministério da Saúde, das secretarias estaduais e municipais de saúde, o local e quantidade de instalação dos equipamentos foi pautada por critérios de necessidade global de radioterapia, de número estimado de casos novos anuais de câncer, de oferta de serviços existentes e de percentuais estaduais de cobertura do sistema de saúde suplementar. Receberão os equipamentos os estados do Alagoas, Bahia, Ceará, Maranhão, Paraíba, Pernambuco, Piauí, Sergipe, Acre, Amapá, Amazonas, Rondônia, Roraima, Tocantins, Goiás, Mato Grosso, Mato Grosso do Sul, Paraná, Rio Grande do Sul, Santa Catarina, São Paulo e Rio de Janeiro, e Distrito Federal (Brasil, 2013).

O Ministério da Saúde tem investido na melhoria do acesso da população a prevenção, exames e tratamentos do câncer, promovendo o investimento em oncologia em 26%, ao aumentar essa cifra de R$ 1,9 bilhão para R$ 2,1 bilhões de 2010 a 2012. A aplicação desse investimento possibilitou ampliar em 17,3% o número de sessões de radioterapia, saltando de 7,6 milhões para mais de nove milhões. No que tange a quimioterapia, o aumento de recursos foi de 14,8%, passando de 2,2 milhões para 2,5 milhões. Em 2013, o orçamento do governo federal foi ampliado de R$ 172,1 milhões em 2012 para R$ 380,3 milhões em 2013, uma elevação de 120% no

orçamento destinado a procedimentos cirúrgicos de oncológica. Nesse período houve também a inclusão de novos tipos de cirurgia oncológica.

A partir de 2009, o governo brasileiro passou a atuar também no sentido de aperfeiçoar a gestão de insumos, que passaram a ser comprados de maneira centralizada pelo MS, com a finalidade de reduzir os custos uma vez que passa a realizar compras em grande escala e não mais de maneira pulverizada e localizada. Por exemplo, ao adquirir o Glivec em grande escala, um medicamento indicado para o tratamento de Leucemia Mieloide Crônica, o governo federal obteve uma redução significativa no preço do medicamento (de mais de 50%) e, com isso, gerou uma economia de aproximadamente R$ 400 milhões no decorrer do período do acordo (de 2010 a 2012). A centralização da compra desse medicamento está na portaria 90 publicada no Diário Oficial da União em 16/03/2011(Brasil, n.d.) (Brasil, 2014). Outra forte atuação aconteceu com a expansão no rol de medicamentos de alto custo ofertados gratuitamente pelo SUS, com a inclusão de drogas biológicas modernas como o mesilato de imatinibe (para o tratamento da leucemia mieloide crônica e tumor do estroma gastrointestinal), o trastuzumabe (para tratar câncer de mama), o l-asparaginase (para tratamento de linfoma linfoblástico) e, agora, o rituximabe (usado no tratamento de linfomas) (Brasil, 2014)(Brasil, 2013).

Em 2014, o SUS incorporou o exame PET-CT (tomografia por emissão de pósitrons) para pacientes com câncer de pulmão, câncer colorretal e de linfoma de Hodgkin e linfoma não Hodgkin (Ministério da Saúde Brasil, 2014*b*) (Ministério da Saúde Brasil and Secretaria de Ciência, Tecnologia e Insumos Estratégicos, 2014*a*) (Ministério da Saúde Brasil and Secretaria de Ciência, Tecnologia e Insumos Estratégicos, 2014*b*). A incorporação do exame ao SUS permite avaliar o grau de avanço do tumor e a extensão da doença no corpo do paciente. O novo exame da PET-CT representa uma melhoria do diagnóstico e tratamento desses tipos de câncer, e permite diminuir os exames e as cirurgias desnecessárias, bem como reduzir a morbidade, a mortalidade e os custos associados ao tratamento dessas doenças. O PET-CT é usado para avaliar viabilidade de cirurgias, pois se o estágio estiver muito avançado, a operação não é recomendável nos casos de câncer de pulmão e para câncer colorretal em pessoas com metástase hepática. No caso de linfomas, o

exame é feito antes e depois da quimioterapia para avaliar a extensão da doença e a resposta ao tratamento.

Outras tecnologias de imagem que são utilizadas para diagnóstico e estadiamento de diversos cânceres no SUS. Entre essas tecnologias são oferecidas radiografia simples, mamografia, cintilografia, ultrassonografia, tomografia computadorizada (CT) e ressonância magnética (MRI)(Ministério da Saúde Brasil Brasil and Portal Brasil, 2014).

6.3 Tempo entre diagnóstico e início de tratamento

Segundo relatório de investigação promovido pelo TCU (Tribunal de Contas da União) os pacientes do SUS esperam, em média, até 70 dias para iniciar o tratamento com quimioterapia após o diagnóstico de câncer; ao passo que para o início do tratamento por radioterapia, a espera é maior e ultrapassa os cem dias. O relatório aponta ainda para o fato de que apenas 35,6% dos pacientes iniciam a quimioterapia após a detecção do tumor. Embora o governo federal esteja promovendo a expansão da cobertura para tratamento contra o câncer, reforçando atenção básica, desde 2005, a época da investigação do TCU, o Brasil contava com 276 serviços cadastrados. Em 2013 esses serviços habilitados em oncologia eram 277, distribuídos da seguinte maneira: 134 no Sudeste, 63 no Sul; 48 no Nordeste, 20 no Centro-Oeste e 12 no Norte. As unidades oferecem radioterapia, quimioterapia e cirurgia oncológica. Dentro dessas unidades de serviço cadastradas para cuidados oncológicos, a quimioterapia respondia por 74,1% dos procedimentos realizados, por outro lado, a radioterapia, com 11,7%, seguida por cirurgias com 9%.

Foram apontados como problema para expansão da cobertura de atendimento oncológico, a falta de equipamentos para radioterapia. As regiões Norte e Nordeste foram apontadas como as regiões com situação mais crítica para atendimentos dessa natureza, e a existência de uma não homogeneidade na distribuição geográfica das unidades de serviço – chamadas “vazios no atendimento" – em todas as demais regiões (TRIBUNAL DE CONTAS DA UNIÃO, 2011).

Segundo dados do Sistema de Informação do Câncer (SISCAN), atualmente, 78% dos pacientes com câncer em estágio inicial recebem tratamento em até 60 dias. Desses, 52% conseguem ser atendidos em 15 dias. Entre os pacientes com

câncer em estágio avançado, 79% recebem tratamento em até 60 dias. Chega a 44% os que conseguem ser atendidos em 15 dias (Secretaria de Estado de Saúde do Distrito Federal, 2014).

Em 2016, no Fórum Estadão Saúde, realizado a 10 de Agosto em São Paulo, o ministro da Saúde apresentou dados do SISCAN que mostravam um aumento de 34% no número de pacientes do SUS para tratar câncer - 292 mil pacientes em 2012, para 393 mil pacientes em 2015. Destes, 57% conseguiram iniciar a terapia dentro do prazo de 60 dias entre o diagnóstico (assinatura do laudo) e primeira intervenção de tratamento (Estadão, 2016*b*; Portal Saúde, MS, 2016).

6.3.1 A lei dos 60 dias para início de tratamento

Considerando a magnitude e importância do câncer para a saúde pública, o governo brasileiro regulamentou a Lei nº 12.732/12 – em vigor desde 23/05/2013 –, a qual estabelece que o primeiro tratamento oncológico no SUS deve se iniciar no prazo máximo de 60 dias a partir da assinatura do laudo patológico ou em prazo menor conforme necessidade terapêutica do caso registrada no prontuário do paciente (Casa Civil Brasil, 2013).

Uma medida do Ministério da Saúde para garantir o cumprimento da lei em todo o Brasil é a realização de visitas aos hospitais que atendem pelo SUS pela Comissão de Monitoramento e Avaliação do cumprimento da Lei nº 12.732, de caráter permanente. Essa comissão tem entre suas atribuições acompanhar os processos de implantação do Sistema de Informação do Câncer (SISCAN) e a execução dos planos regionais de oncologia. Nesses hospitais também devem ser monitoradas as condições de funcionamento e a capacidade de ofertar o atendimento oncológico com agilidade (Ministério da Saúde Brasil and Gabinete do Ministro, 2013). A Portaria GM nº 876 publicada pelo Ministério da Saúde regulamenta o SISCAN como ferramenta oficial para gerenciar o cumprimento do prazo estabelecido. O SISCAN possui um módulo denominado “Tempo Diagnóstico/Tratamento”, que permite o registro de todos os diagnósticos confirmados de câncer, bem como, data e modalidade do primeiro tratamento realizado. Essa informação tornou-se obrigatória apenas em 22 de maio de 2013, data em que a Portaria GM nº 876 entrou em vigor (Instituto Nacional de Câncer

José Alencar Gomes da Silva, 2013; Ministério da Saúde Brasil and Gabinete do Ministro, 2013).

7. Bases de dados do SUS

O SUS dispõe hoje de sistemas públicos que permitem a obtenção de informações epidemiológicas e sociodemográficas para subsidiar diversas esferas de gestão pública na definição de prioridades que visem à prevenção e controle de doenças. Entre essas bases podemos citar o Sistema de Informações Hospitalares (SIH-SUS), Sistema de Informações Ambulatoriais (SIA-SUS) e Sistema de Informações sobre Mortalidade (SIM).

Apesar das limitações de qualidade e abrangência, o SIH-SUS e o SIA-SUS, operam como a única fonte de informação sobre internações hospitalares e atendimentos ambulatoriais para grande parte dos estados e municípios do Brasil. Originalmente esses sistemas tem como finalidade controle administrativo para faturamento, projeção e controle de provisão e gastos com a assistência ambulatorial e hospitalar. Nos últimos anos, eles passaram a auxiliar no controle de provisão e gastos com a assistência ambulatorial e hospitalar ao terem suas informações reajustadas para análise em pesquisa (Scatena and Tanaka, 2001). A conversão dessas bases de dados administrativas pode ser explorada no sentido de subsidiar a avaliação dos serviços do SUS, quer do ponto de vista do financiamento da assistência à saúde, como da produção e resolubilidade dos serviços de saúde. Apresentamos mais detalhadamente essas bases de dados na seção seguinte.

7.1 Sistema de Informações Ambulatoriais do SUS - SIA/SUS (APAC)

O Sistema de Informações Ambulatoriais do SUS (SIA/SUS) foi criado em 1992 e implantado a partir de julho de 1994, nas Secretarias Estaduais, as quais substituíam os sistemas controladores do financiamento de atendimentos ambulatoriais GAP e SICAPS. Em 1996, aconteceu um movimento para ampla implementação do novo sistema nas Secretarias Municipais de Saúde – então chamadas de gestão semiplenas – pela NOB 96. Em 1997, esse sistema passou a processar além dos tradicionais BPA (Boletim de Produção Ambulatorial), um documento numerado e autorizado chamado Autorização de Procedimento de Alta

Complexidade “APAC”(Instituto Brasileiro de Geografia e Estatística IBGE Brasil, n.d.). A APAC é inaugurada no SUS com a autorização para procedimentos dialíticos (APAC – Terapia Renal Substitutiva – TRS) em 1998 (Ministério da Saúde Brasil, 2013*d*). A organização da base de dados de APAC se consolida em 1999.

Em janeiro de 2008, o BPA foi divido em dois, o BPA-I – Boletim de Produção Ambulatorial Individualizado; e BPA-C – Boletim de Produção Ambulatorial Consolidado. O BPA – Consolidado contempla procedimentos registrados de forma agregada por CBO do profissional, idade do paciente em uma competência de atendimento de um estabelecimento (CNES). Ao passo que o BPA – Individualizado se orienta para procedimentos registrados pelo CNS/CBO do profissional, informação individualizada do Paciente (CNS, Nome, Data Nascimento, CID e Caráter de Atendimento, Sexo e Município de Residência).

Desde a sua criação, em plataforma 16 bits, os dados do SIA-SUS são gerenciados em Clipper com DBF, disseminados em arquivos DBC – um DBF comprimido(Instituto Brasileiro de Geografia e Estatística IBGE Brasil, n.d.).

O SIA-SUS é o sistema que permite aos gestores locais o processamento das informações de atendimento ambulatorial (não hospitalar) registrados nos aplicativos de captação do atendimento ambulatorial pelos prestadores públicos e privados contratados/conveniados pelo SUS. Esse sistema é alimentado pela transcrição de produção dos documentos BPA (Boletim de Produção Ambulatorial) e APAC (Autorização de Procedimentos de Alta Complexidade. Essas informações são posteriormente consolidadas e validadas. A validação acontece contra parâmetros orçamentários estipulados pelo Gestor de saúde, antes de aprovar o pagamento. Uma base de dados com a totalidade dos procedimentos realizados é enviada ao DATASUS pelos gestores. Essa base de dados guarda os valores devidos a rede de estabelecimentos do Gestor de saúde.

O DATASUS gera arquivos para tabulação (Tabnet/Tabwin) contendo estes atendimentos com periodicidade mensal. Incorporadas às informações do sistema SIH/SUS, o SIA/SUS fornece ao SAS/DRAC os valores do Teto de Financiamento a serem repassados para os gestores(Instituto Brasileiro de Geografia e Estatística IBGE Brasil, n.d.). Portanto as informações do SIA/SUS constituem um

importante instrumento de gestão, apoiando as ações de planejamento, programação, regulação, avaliação, controle e auditoria da assistência ambulatorial.

Os dois principais objetos de captação da informação para o sistema são o BPA (Boletim de Produção Ambulatorial) e a APAC (Autorização de Procedimentos de Alta Complexidade). O BPA é preenchido pelo estabelecimento de saúde que realiza o procedimento e, então, seus dados são processados de forma descentralizadas no nível municipal e estadual, para serem finalmente consolidados no nível nacional. As APAC concentram as informações dos procedimentos de alta complexidade e são armazenados por visita do paciente. Os arquivos de APAC são compostos conforme o Tipo de Laudo (Tipo de Atendimento) de APAC. Os laudos de APAC são os seguintes: Oncologia (AQ – quimioterapia, AR – radioterapia), Nefrologia (AN), acompanhamento a Cirurgia bariátrica (AB), Medicamentos (AM) e Laudos diversos (AD)(Ministério da Saúde BRASIL, Secretaria Executiva DATASUS, and GEDINF – Gerência de Disseminação de Informações em Saúde, 2010).

Sobre a organização dos registros na base de dados disseminada, cabe ressaltar que o instrumento APAC gera diversos registros no arquivo de disseminação, ou seja, há um registro para cada código de procedimento realizado na APAC – seja o procedimento principal ou os procedimentos secundários. Nos arquivos de APAC, o procedimento se refere ao procedimento principal. Dessa forma, cada paciente pode apresentar vários números de APAC e diversos registros de procedimentos para um mesmo número de APAC.

As informações a seguir (TABELA 1 e TABELA 2) se referem aos atendimentos ambulatoriais em pacientes que realizaram APAC nos meses a partir de janeiro de 2008, quando foi implantada a Tabela de Procedimentos, Medicamentos, Órteses e Próteses e Materiais Especiais do Sistema Único de Saúde – SUS, instituída pela portaria GM/MS n.º 321 de 08 de fevereiro de 2007. A partir de janeiro de 2008 o identificador de um paciente (chave primária) passou a ser obrigatoriamente o Cartão Nacional de Saúde (CNS)(MINISTÉRIO DA SAÚDE - Manual_Operacional_APAC_v_1_1.pdf [Internet], n.d.; Ministério da Saúde BRASIL, Secretaria Executiva DATASUS, and GEDINF – Gerência de Disseminação de Informações em Saúde, 2010).

As principais variáveis captadas pelo SIA/SUS para BPA – Boletim de Produção Ambulatorial são 1) identificadores de estabelecimento: código, município, regime jurídico; 2) dados de procedimentos: código do procedimento, código do estabelecimento ou profissional, tipo do profissional (médico, enfermeiro, medicina ocupacional, etc.), programa de saúde (diabete, hipertensão, medicina ocupacional, materno-infantil, etc.), Tipo do atendimento (emergência, vacinação, etc.), Custo do procedimento; 3) identificadores do paciente – sexo, idade, município de residência, município de nascimento.

Para APAC são os seguintes grupos de informação: 1) identificadores de estabelecimento – código, município, regime jurídico; 2) dados de procedimentos: data, município, grupo (hemodiálise, quimioterapia, cirurgia ambulatorial, patologia clínica, etc.), principal procedimento, número da APAC, código da doença principal (CID), custo; 3) identificadores do paciente – sexo, idade, município de residência, município de nascimento.

TABELA 1 – Layout dos arquivos de APAC de Quimioterapia (AQ)

SEQUÊNCIA	CAMPO	TIPO	TAMANHO	DESCRIÇÃO
1	AP_MVM	CHAR	6	Data de Processamento / Movimento (AAAAMM)
2	AP_CONDIC	CHAR	2	Sigla do Tipo de Gestão que o Estado ou Município está habilitado
3	AP_GESTAO	CHAR	6	Unidade de Federação + Código Município de Gestão ou UF0000 se a Unidade está sob Gestão Estadual
4	AP_CODUNI	CHAR	7	Código do CNES do Estabelecimento de Saúde
5	AP_AUTORIZ	CHAR	13	Número da APAC
6	AP_CMP	CHAR	6	Data de Atendimento ao Paciente / Competência (AAAAMM)
7	AP_PRIPAL	CHAR	10	Procedimento Principal da APAC
8	AP_VL_AP	NUMERIC	(12.2)	Valor Total da APAC Aprovado
9	AP_UFMUN	CHAR	6	Unidade da Federação + Município do Estabelecimento
10	AP_TPUPS	CHAR	2	Tipo de Estabelecimento

SEQUÊNCIA	CAMPO	TIPO	TAMANHO	DESCRIÇÃO
11	AP_TIPPRE	CHAR	2	Tipo de Prestador
12	AP_MN_IND	CHAR	1	Mantida / Individual
13	AP_CNPJCPF	CHAR	14	CNPJ do Estabelecimento executante
14	AP_CNPJMNT	CHAR	14	CNPJ MANTENEDORA
15	AP_CNSPCN	CHAR	15	CNS do Paciente
16	AP_COIDADE	CHAR	3	Código da Idade
17	AP_NUIDADE	CHAR	2	Numero da Idade
18	AP_SEXO	CHAR	1	Sexo
19	AP_RACACOR	CHAR	2	Cor / Raça
20	AP_MUNPCN	CHAR	6	UF + Município de Residência do paciente
21	AP_UFNACIO	CHAR	3	Nacionalidade do paciente
22	AP_CEPPCN	CHAR	8	CEP do paciente
23	AP_UFDIF	CHAR	2	Indica se a UF de residência do paciente é diferente da UF de localização do estabelecimento (N=não, S=sim)
24	AP_MNDIF	CHAR	2	Indica se o município de residência do paciente é diferente do município de localização do estabelecimento (N=não, S=sim)
25	AP_DTINIC	CHAR	8	Data de INÍCIO validade
26	AP_DTFIM	CHAR	8	Data de FIM validade
27	AP_TPATEN	CHAR	2	Tipo de Atendimento de APAC
28	AP_TPAPAC	CHAR	1	Indica se a APAC é 1 – inicial, 2 – continuidade, 3 – única
29	AP_MOTSAI	CHAR	2	Motivo de Saída e Permanência
30	AP_OBITO	CHAR	1	Indicador de Óbito
31	AP_ENCERR	CHAR	1	Indicador Encerramento
32	AP_PERMAN	CHAR	1	Indicador Permanência
33	AP_ALTA	CHAR	1	Indicador de Alta
34	AP_TRANSF	CHAR	1	Indicar de Transferência
35	AP_DTOCOR	CHAR	8	Data de Ocorrência que substitui a data de FIM de validade
36	AP_CODEMI	CHAR	10	Código do Órgão emissor

SEQUÊNCIA	CAMPO	TIPO	TAMANHO	DESCRIÇÃO
37	AP_CATEND	CHAR	2	Caráter do Atendimento
38	AP_APACANT	CHAR	13	Número APAC Anterior
39	AP_UNISOL	CHAR	7	Código CNES do Estabelecimento Solicitante
40	AP_DTSOLIC	CHAR	8	Data da Solicitação
41	AP_DTAUT	CHAR	8	Data da Autorização
42	AP_CIDCAS	CHAR	4	CID Causas Associadas
43	AP_CIDPRI	CHAR	4	CID Principal
44	AP_CIDSEC	CHAR	4	CID Secundário
45	AQ_CID10	CHAR	4	CID 10 Topografia
46	AQ_LINFIN	CHAR	1	Linfonodos regionais invadidos (S = Sim; N =Não; 3= Não Avaliáveis)
47	AQ_ESTADI	CHAR	1	Estádio – UICC (0;1;2;3;4)
48	AQ_GRAHIS	CHAR	2	Grau Histopatológico
49	AQ_DTIDEN	CHAR	8	Data da identificação patológica do caso (AAAAMMDD)
50	AQ_TRANTE	CHAR	1	Tratamentos anteriores (S=Sim; N=Não)
51	AQ_CIDINI1	CHAR	4	CID 1o. Tratamento anterior
52	AQ_DTINI1	CHAR	8	Data de inicio (AAAAMMDD) 1º tratamento anterior
53	AQ_CIDINI2	CHAR	4	CID 2o. Tratamento anterior
54	AQ_DTINI2	CHAR	8	Data de inicio (AAAAMMDD) 2º tratamento anterior
55	AQ_CIDINI3	CHAR	4	CID 3o. Tratamento anterior
56	AQ_DTINI3	CHAR	8	Data de inicio (AAAAMMDD) 3º tratamento anterior
57	AQ_CONTTR	CHAR	1	Continuidade do tratamento (S=Sim; N=Não)
58	AQ_DTINTR	CHAR	8	Data de INICIO do tratamento solicitado (AAAAMMDD)
59	AQ_ESQU_P1	CHAR	5	ESQUEMA (Sigla ou abrev) – 5 primeiras posições
60	AQ_TOTMPL	CHAR	3	Total de MESES Planejados

SEQUÊNCIA	CAMPO	TIPO	TAMANHO	DESCRIÇÃO
61	AQ_TOTMAU	CHAR	3	Total de MESES Autorizados
62	AQ_ESQU_P2	CHAR	10	ESQUEMA (Sigla ou abrev) - 10 últimas posições

TABELA 2 - Layout dos arquivos de APAC de Radioterapia (AR)

SEQUÊNCIA	CAMPO	TIPO	TAMANHO	DESCRIÇÃO
1	AP_MVM	CHAR	6	Data de Processamento / Movimento (AAAAMM)
2	AP_CONDIC	CHAR	2	Sigla do Tipo de Gestão que o Estado ou Município está habilitado
3	AP_GESTAO	CHAR	6	Unidade de Federação + Código Município de Gestão ou UF0000 se a Unidade está sob Gestão Estadual
4	AP_CODUNI	CHAR	7	Código do CNES do Estabelecimento de Saúde
5	AP_AUTORIZ	CHAR	13	Número da APAC
6	AP_CMP	CHAR	6	Data de Atendimento ao Paciente / Competência (AAAAMM)
7	AP_PRIPAL	CHAR	10	Procedimento Principal da APAC
8	AP_VL_AP	NUMERIC	20.2	Valor Total da APAC Aprovado
9	AP_UFMUN	CHAR	6	Unidade da Federação + Município do Estabelecimento
10	AP_TPUPS	CHAR	2	Tipo de Estabelecimento
11	AP_TIPPRE	CHAR	2	Tipo de Prestador
12	AP_MN_IND	CHAR	1	Mantida / Individual
13	AP_CNPJCPF	CHAR	14	CNPJ do Estabelecimento executante
14	AP_CNPJMNT	CHAR	14	CNPJ MANTENEDORA
15	AP_CNSPCN	CHAR	15	CNS do Paciente
16	AP_COIDADE	CHAR	3	Código da Idade
17	AP_NUIDADE	CHAR	2	Numero da Idade
18	AP_SEXO	CHAR	1	Sexo
19	AP_RACACOR	CHAR	2	Cor / Raça
20	AP_MUNPCN	CHAR	6	UF + Município de Residência do paciente

SEQUÊNCIA	CAMPO	TIPO	TAMANHO	DESCRIÇÃO
21	AP_UFNACIO	CHAR	3	Nacionalidade do paciente
22	AP_CEPPCN	CHAR	8	CEP do paciente
23	AP_UFDIF	CHAR	2	Indica se a UF de residência do paciente é diferente da UF de localização do estabelecimento (N=não, S=sim)
24	AP_MNDIF	CHAR	2	Indica se o município de residência do paciente é diferente do município de localização do estabelecimento (N=não, S=sim)
25	AP_DTINIC	CHAR	8	Data de INÍCIO validade
26	AP_DTFIM	CHAR	8	Data de FIM validade
27	AP_TPATEN	CHAR	2	Tipo de Atendimento de APAC
28	AP_TPAPAC	CHAR	1	Indica se a APAC é 1 – inicial, 2 – continuidade, 3 – única
29	AP_MOTSAI	CHAR	2	Motivo de Saída e Permanência
30	AP_OBITO	CHAR	1	Indicador de Óbito
31	AP_ENCERR	CHAR	1	Indicador Encerramento
32	AP_PERMAN	CHAR	1	Indicador Permanência
33	AP_ALTA	CHAR	1	Indicador de Alta
34	AP_TRANSF	CHAR	1	Indicar de Transferência
35	AP_DTOCOR	CHAR	8	Data de Ocorrência que substitui a data de FIM de validade
36	AP_CODEMI	CHAR	10	Código do Órgão emissor
37	AP_CATEND	CHAR	2	Caráter do Atendimento
38	AP_APACANT	CHAR	13	Número APAC Anterior
39	AP_UNISOL	CHAR	7	Código CNES do Estabelecimento Solicitante
40	AP_DTSOLIC	CHAR	8	Data da Solicitação
41	AP_DTAUT	CHAR	8	Data da Autorização
42	AP_CIDCAS	CHAR	4	CID Causas Associadas
43	AP_CIDPRI	CHAR	4	CID Principal
44	AP_CIDSEC	CHAR	4	CID Secundário
45	AR_SMRD	CHAR	3	
46	AR_CID10	CHAR	4	CID 10 Topografia

SEQUÊNCIA	CAMPO	TIPO	TAMANHO	DESCRIÇÃO
47	AR_LINFIN	CHAR	1	Linfonodos regionais invadidos (S = Sim; N =Não; 3= Não Avaliáveis)
48	AR_ESTADI	CHAR	1	Estádio – UICC (0;1;2;3;4)
49	AR_GRAHIS	CHAR	2	Grau Histopatológico
50	AR_DTIDEN	CHAR	8	Data da identificação patológica do caso (AAAAMMDD)
51	AR_TRANTE	CHAR	1	Tratamentos anteriores (S=Sim; N=Não)
52	AR_CIDINI1	CHAR	4	CID 1o. Tratamento anterior
53	AR_DTINI1	CHAR	8	Data de INICIO (AAAAMMDD) 1º tratamento anterior
54	AR_CIDINI2	CHAR	4	CID 2o. Tratamento anterior
55	AR_DTINI2	CHAR	8	Data de INICIO (AAAAMMDD) 2º tratamento anterior
56	AR_CIDINI3	CHAR	4	CID 3o. Tratamento anterior
57	AR_DTINI3	CHAR	8	Data de inicio (AAAAMMDD) 3º tratamento anterior
58	AR_CONTTR	CHAR	1	Continuidade do tratamento (S=Sim; N=Não)
59	AR_DTINTR	CHAR	8	Data de INICIO do tratamento solicitado (AAAAMMDD)
60	AR_FINALI	CHAR	1	Finalidade Do Tratamento (1=RADICAL; 2=ADJUVANTE; 3=ANTIÁLGICA; 4=PALIATIVA; 5=PRÉVIA;6=ANTIHEMORRÁGICA)
61	AR_CIDTR1	CHAR	4	CID Topográfico 1º
62	AR_CIDTR2	CHAR	4	CID Topográfico 2º
63	AR_CIDTR3	CHAR	4	CID Topográfico 3º
64	AR_NUMC1	CHAR	3	No. Campo/Inserções 1º
65	AR_INIAR1	CHAR	8	Data de INICIO 1º (AAAAMMDD)
66	AR_INIAR2	CHAR	8	Data de INICIO 2º (AAAAMMDD)
67	AR_INIAR3	CHAR	8	Data de INICIO 3º (AAAAMMDD)
68	AR_FIMAR1	CHAR	8	Data de Fim 1º (AAAAMMDD)
69	AR_FIMAR2	CHAR	8	Data de FIM 2º (AAAAMMDD)
70	AR_FIMAR3	CHAR	8	Data de Fim 3º (AAAAMMDD)
71	AR_NUMC2	CHAR	3	No. Campo/Inserções 2º
72	AR_NUMC3	CHAR	3	No. Campo/Inserções 3º

7.2 Sistema de Informações Hospitalares do SUS (SIH-SUS)

O Sistema de Informações Hospitalares do SUS (SIH-SUS) surgiu sob a responsabilidade do Ministério da Previdência e Assistência Social com o nome de Sistema de Assistência Médico-Hospitalar da Previdência Social (SAMHPS) – com uma lógica predominantemente contábil, captando variáveis essencialmente dessa natureza (Scatena and Tanaka, 2001). Em 1981 o Sistema de Autorização Hospitalar substituiu o sistema GIH (Guia de Internação Hospitalar), sendo ainda um sistema de processamento centralizado em mainframes (UNISYS e ABC-BULL). Em 1992, o SIH-SUS se tornou o primeiro sistema do DATASUS a ter captação de dados feita por microcomputadores (AIH em DISQUETE) e realizada na origem da geração dos dados – isto é, os dados eram captados nos próprios usuários, dando fim ao modelo de polos de digitação. Até abri de 2006, o processamento das autorizações para internação hospitalar foi centralizado. A partir desta data, o processamento passou a ser tarefa dos gestores de Secretaria de Saúde através de uma usando plataforma Windows, SGBD Firebird e Linguagem de programação Delphi. Hoje esse sistema se encontra controlado por SGBD Oracle(DATASUS Brasil, n.d.).

O SIH-SUS tem a finalidade de captar, validar e armazenar informações de atendimentos provenientes de internações hospitalares que foram financiadas pelo SUS. As informações que constituem o sistema, começa com o preenchimento da Autorização de Internação Hospitalar – documento operacional do sistema – pelo hospital após a alta hospitalar. Os dados são enviados eletronicamente, dependendo do nível de gestão municipal, para a Secretaria de Saúde municipal ou estadual. A consolidação dos dados é executada a nível nacional pelo DATASUS. Após o processamento dos registros, são gerados relatórios para que os gestores possam efetuar pagamentos dos estabelecimentos de saúde vinculados com o SUS. Todos os dados captados e enviados à União, alimentando assim uma base de dados de todas as internações autorizadas, aprovadas ou não para pagamento, em suas variadas formas de contrato de gestão – para que possam ser repassados os recursos às Secretarias de Saúde (valores de Produção de Média e Alta complexidade, de CNRAC, FAEC e de Hospitais Universitários).

A divulgação dos dados do SIH-SUS, a nível nacional, é realizada mensalmente pelo DATASUS. As principais informações captadas pela AIH, organizadas por natureza, são as seguintes: 1) código, município, regime jurídico do hospital; 2) identificação do paciente – sexo, data de nascimento, idade, município de residência, código postal, ocupação, atividade econômica, nacionalidade; 3) caracterização da hospitalização, com especialidade, tipo de admissão (emergência, eletiva, etc.), data da admissão, data da alta, dias de permanência, tipo e número de dias na UTI, número de dias do acompanhante, motivo da alta, procedimentos realizados, diagnóstico primário e secundário; 4) custo hospitalar – total, serviços hospitalares, serviços profissionais, serviços de diagnóstico e terapia, cuidados neonatais, acompanhante, ortopedia e prótese, sangue, transplante, analgesia obstétrica, UTI; 5) natureza dos procedimentos – código do procedimento, código do estabelecimento ou profissional, tipo do estabelecimento ou profissional (laboratório do hospital, laboratório externo, profissional do hospital, profissional externo), atividade do estabelecimento ou profissional (anestesia, ortopedia, etc.), número de procedimentos, custo. Adicionalmente, dependendo da natureza da hospitalização, podem ser captadas informações sobre gravidez de alto-risco, cuidados pré-natais, infecção hospitalar. Os casos de esterilização têm recolhidos, além das informações padrão, dados sobre número de filhos, escolaridade e métodos de controle de natalidade (Scatena and Tanaka, 2001; Boing, d'Orsi, and Reibnitz, 2010; DATASUS Brasil, n.d.).

TABELA 3 - Layout dos arquivos de AIH reduzida

NOME	TIPO	TAM	DEC	DESCRIÇÃO
UF_ZI	C	6	0	MUNICIPIO GESTOR
ANO_CMPT	C	4	0	ANO COMPETÊNCIA
MES_CMPT	C	2	0	MÊS COMPETÊNCIA
ESPEC	C	2	0	CODIGO DA ESPECIALIDADE
CGC_HOSP	C	14	0	CGC DO HOSPITAL
N_AIH	C	13	0	NÚMERO DA AIH
IDENT	C	1	0	IDENTIFICACAO DO TIPO DE AIH
CEP	C	8	0	CEP DO PACIENTE
MUNIC_RES	C	6	0	CODIGO IBGE DO MUNICIPIO DO PACIENTE
NASC	C	8	0	DATA DE NASCIMENTO DO PACIENTE (AAAAMMDD)

NOME	TIPO	TAM	DEC	DESCRIÇÃO
SEXO	C	1	0	SEXO DO PACIENTE ONDE:1=MASCULINO; 3=FEMININO
UTI_MES_IN	N	2	0	QTD DE DIAS DE UTI NO MES INICIAL DA INTERNACAO
UTI_MES_AN	N	2	0	QTD DE DIAS DE UTI NO MES ANTERIOR A ALTA
UTI_MES_AL	N	2	0	QTD DE DIAS DE UTI NO MES DA ALTA
UTI_MES_TO	N	2	0	QUANTIDADE DE ATOS - UTI
MARCA_UTI	C	2	0	MARCA DE UTI - DEPENDE DO ATO
UTI_INT_IN	N	2	0	DIAS EM UTI NO MÊS INICIAL (tipo ato = 34 ou 56)
UTI_INT_AN	N	2	0	DIAS EM UTI NO MÊS ANTERIOR À ALTA (tipo ato = 34 ou 56)
UTI_INT_AL	N	2	0	DIAS EM UTI NO MÊS DA ALTA (tipo ato = 34 ou 56)
UTI_INT_TO	N	2	0	TOTALIZA DIAS EM UTI (tipo ato = 34 ou 56)
DIAR_ACOM	N	2	0	QTD DE DIARIAS DE ACOMPANHANTE
PROC_SOLIC	C	10	0	CODIGO DO PROCEDIMENTO SOLICITADO
PROC_REA	C	10	0	CODIGO DO PROCEDIMENTO REALIZADO
VAL_SH	N	13	2	VALOR DOS SERVIÇOS HOSPITALARES
VAL_SP	N	13	2	VALOR DOS SERVIÇOS PRESTADOS POR TERCEIROS
VAL_SADT	N	13	2	VALOR DE SADT
VAL_RN	N	13	2	VALOR DE RECEM-NATO
VAL_ACOMP	N	13	2	VALOR DE ACOMPANHANTE
VAL_ORTP	N	13	2	VALOR DE ÓRTESE E PRÓTESE
VAL_SANGUE	N	13	2	VALOR DE SANGUE
VAL_SADTSR	N	11	2	VALOR DE SADT SEM RATEIO
VAL_TRANSP	N	13	2	VALOR DE TRANSPLANTE
VAL_OBSANG	N	11	2	VALOR DE ANESTESIA
VAL_PED1AC	N	11	2	VALOR DE PEDIATRIA
VAL_TOT	N	14	2	VALOR TOTAL
VAL_UTI	N	8	2	VALOR DE UTI
US_TOT	N	10	2	VALOR TOTAL EM DOLAR
DT_INTER	C	8	0	DATA DA INTERNACAO (AAAAMMDD)
DT_SAIDA	C	8	0	DATA DA SAIDA (AAAAMMDD)
DIAG_PRINC	C	4	0	CODIGO DO DIAGNOSTICO PRINCIPAL
DIAG_SECUN	C	4	0	CODIGO DO DIAGNOSTICO SECUNDARIO
COBRANCA	C	2	0	MOTIVO DE COBRANCA
NATUREZA	C	2	0	NATUREZA JURÍDICA DO HOSPITAL
GESTAO	C	1	0	CODIGO DO ORGAO EMISSOR DO GESTOR (E=2,outros=1)
RUBRICA	N	5	0	NUMERO DA RUBRICA

NOME	TIPO	TAM	DEC	DESCRIÇÃO
IND_VDRL	C	1	0	INDICA EXAME VDRL (SIM=1)
MUNIC_MOV	C	6	0	MUNICIPIO DO HOSPITAL
COD_IDADE	C	1	0	CODIGO DA IDADE
IDADE	N	2	0	IDADE
DIAS_PERM	N	5	0	DIAS DE PERMANÊNCIA
MORTE	N	1	0	INDICA ÓBITO (SIM=1)
NACIONAL	C	2	0	CODIGO DA NACIONALIDADE DO PACIENTE
NUM_PROC	C	4	0	NUMERO DO PROCESSAMENTO
CAR_INT	C	2	0	CARATER DE INTERNACAO
TOT_PT_SP	N	6	0	TOTAL DE PONTOS EM SP
CPF_AUT	C	11	0	CPF DO AUDITOR QUE AUTORIZOU O HOMONIMO
HOMONIMO	C	1	0	MARCA DE HOMONIMO (0 = não , 1 = sim)
NUM_FILHOS	N	2	0	QTD DE FILHOS DO PACIENTE
INSTRU	C	1	0	GRAU DE INSTRUCAO
CID_NOTIF	C	4	0	CID DE NOTIFICACAO
CONTRACEP1	C	2	0	METODO CONTRACEPTIVO
CONTRACEP2	C	2	0	METODO CONTRACEPTIVO
GESTRISCO	C	1	0	GESTANTE DE ALTO RISCO (0=NAO; 1=SIM)
INSC_PN	C	10	0	NUMERO DA GESTANTE NO PRE-NATAL
SEQ_AIH5	C	3	0	SEQUENCIAL DA AIH TIPO 5 (LONGA PERMANENCIA)
CBOR	C	3	0	CODIGO DO CBO (CODIGO BRASILEIRO DE OCUPAÇÕES)
CNAER	C	3	0	CODIGO DE ACIDENTE DE TRABALHO
VINCPREV	C	1	0	VINCULO COM A PREVIDENCIA
GESTOR_COD	C	3	0	CODIGO DA AUTORIZACAO PAGAMENTO
GESTOR_CPF	C	11	0	NUMERO DO CPF DO GESTOR
GESTOR_DT	C	8	0	DATA DA AUTORIZACAO DADA PELO GESTOR (AAAAMMDD)
CNES	C	7	0	CODIGO DO CNES
INFEHOSP	C	1	0	STATUS DE INFECCAO HOSPITALAR (ONDE: 0=NAO; 1=SIM)
CID_ASSO	C	4	0	CID CAUSA
CID_MORTE	C	4	0	CID DA MORTE

7.3 Sistema de Informação sobre Mortalidade (SIM)

O Sistema de Informação sobre Mortalidade (SIM) foi criado em 1975, visando à obtenção de dados de mortalidade de forma regular e abrangente no

Brasil através do preenchimento das Declarações de Óbito (Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Análise de Situação de Saúde. Brasil, n.d.). Esse sistema tem por finalidade concentrar formalmente informações sobre óbitos ocorridos no Brasil. A base de dados é formada por variáveis – quantitativas e qualitativas sobre óbitos - que permitem, a partir da causa mortis atestada pelo médico, construir indicadores e processar análises epidemiológicas que contribuam para a eficiência da gestão em saúde.

O SIM foi informatizado em 1979 e teve promovida sua implementação gradualmente desde 1994 pelo Ministério da Saúde em todas as unidades da federação(Instituto Brasileiro de Geografia e Estatística BRASIL, n.d.).

Entre as funcionalidades disponíveis no SIM, temos a declaração de óbito informatizada; a geração de arquivos de dados em várias extensões para analises em outros aplicativos; a retroalimentação das informações ocorridas em municípios diferentes da residência do paciente; transmissão de dados automatizada utilizando a ferramenta SISNET – gerando a tramitação dos dados de forma ágil e segura entre os níveis municipal, estadual e federal, bem como um sistema de backup on-line desses níveis do estado brasileiro(Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Análise de Situação de Saúde. Brasil, n.d.; Ministério da Saúde Brasil and Fundação Nacional de Saúde Brasil, 2001).

Em um sistema de mortalidade, as unidades de investigação e análise são os óbitos. Assim, a captação dessa informação é dada pelo Declaração de Óbito (DO) – a qual é padronizada e distribuída, em três vias, para todo o país pelo Ministério da Saúde. Em 1999 foi adotado o atual modelo de DO, acompanhado de um novo sistema informatizado para a sua manipulação(Ministério da Saúde Brasil and Fundação Nacional de Saúde Brasil, 2001). A DO deve ser preenchida pelo médico que realizou o atendimento, ou por duas pessoas qualificadas que tenham presenciado ou verificado a morte – isso no caso da ausência desse médico. O recebimento das DO, para alimentação do sistema de mortalidade, é feito pela Secretaria de Saúde do Município ou do Estado no Estabelecimento de Saúde.

As principais variáveis captadas pelo SIM são as seguintes: 1) certificado de óbito: número de registro e tipo de óbito (fetal /não fetal); 2) identificação do falecido: local de nascimento, data de nascimento, idade, sexo,

raça/cor, estado civil, escolaridade, ocupação, distrito e município de residência; 3) identificação do óbito: data, local, estabelecimento de saúde, município; 4) caraterização do óbito: Causa primária da morte (CID), causas secundárias, prováveis circunstâncias de morte não natural. Os óbitos enquadrados em caso de morte fetal ou de criança com menos de 1 ano recebem tratamento diferenciado. Nessas situações, é agregada ao registro de óbito informações sobre a mãe – idade, escolaridade, ocupação, número de filhos nascidos vivos, número de filhos nascidos mortos, tipo de gravidez (única ou múltipla), semanas de gestação, tipo de parto (vaginal ou cesáreo), doenças relativas ao parto, peso ao nascer(Instituto Brasileiro de Geografia e Estatística IBGE Brasil, n.d.; Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Análise de Situação de Saúde. Brasil, n.d.).

Os benefícios da existência de um sistema de informação de mortalidade, como implementado no Brasil, são a produção de estatísticas de mortalidade, construção dos principais indicadores de saúde, análises estatísticas, epidemiológicas e sócio-demográficas.

TABELA 4 – Estrutura do arquivo de Mortalidade 2006-2012.

SEQ	NOME	TIPO/TAM	DESCRIÇÃO
1	NUMERODO	C(08)	Número da DO, sequencial por UF informante e por ano.
2	TIPOBITO	C(01)	1- óbito fetal 2- óbito não fetal
3	DTOBITO	C(08)	Data do óbito, no formato ddmmaaaa .
4	HORAOBITO	C(05)	Hora do falecimento.
5	NATURAL	C(03)	Naturalidade, conforme a tabela de países. Se for brasileiro, porém, o primeiro dígito contém 8 e os demais
6	DTNASC	C(08)	Data de nascimento do falecido.
6	DTNASC	C(08)	Data de nascimento do falecido.
7	IDADE	C(03)	Idade, composto de dois subcampos. O primeiro, de 1 dígito, indica a unidade da idade, conforme a tabela
8	SEXO	C(01)	Sexo, conforme a tabela- 0- Ignorado 1- Masculino 2- Feminino
9	RACACOR	C(01)	Raça/Cor- 1-Branca 2-Preta 3-Amarela 4-Parda 5- Indígena
10	ESTCIVIL	C(01)	Estado civil, conforme a tabela- 1- Solteiro 2-Casado 3- Viúvo 4- Separado judicialmente 9-Ignorado

11	ESC	C(01)	Escolaridade, Anos de estudo concluídos- 1- Nenhuma 2- 1 a 3 anos 3- 4 a 7 anos 4-
12	OCUP	C(06)	Ocupação, conforme a Classificação Brasileira de Ocupações (CBO-2002)
13	CODMUNRES	C(07)	Município de residência do falecido
14	CODBAIRES	C(08)	Código do Bairro de residência
15	LOCOCOR	C(01)	Local de ocorrência do óbito, conforme a tabela- 9- Ignorado 1- Hospital 2- Outro estab saúde 3- Domicílio
16	CODESTAB	C(07)	Código do estabelecimento.
17	CODMUNOCOR	C(07)	Município de ocorrência do óbito, conforme códigos IBGE.
18	CODBAIOCOR	C(08)	Código do bairro de ocorrência.
19	IDADEMAE	C(02)	Idade da mãe em anos.
20	ESCMAE	C(01)	Escolaridade, Anos de estudo concluídos- 1- Nenhuma 2- 1 a 3 anos 3- 4 a 7 anos 4-
21	OCUPMAE	C(05)	Ocupação da mãe, conforme codificação de OCUPACAO
22	QTDFILVIVO	C(02)	Número de filhos vivos.
23	QTDFILMORT	C(02)	Número de filhos mortos, ignorados, não incluindo o próprio.
24	GRAVIDEZ	C(01)	Tipo de gravidez, conforme a tabela- 9- Ignorado 1- Única 2- Dupla 3- Tripla e mais
25	GESTACAO	C(01)	Semanas de gestação, conforme as tabelas- 9- Ignorado 1- Menos de 22 semanas 2- 22 a 27 semanas
26	PARTO	C(01)	Tipo de parto, conforme a tabela- 9- Ignorado 1- Vaginal 2- Cesáreo
27	OBITOPARTO	C(01)	Morte em relação ao parto, conforme tabela- 9- Ignorado 1- Antes 2- Durante 3- Depois
28	PESO	C(04)	Peso ao nascer, em gramas.
29	NUMERODN	C(08)	Número da DN
30	OBITOGRAV	C(01)	Morte durante a Gravidez conforme tabela- 9- Ignorado 1- Sim 2- Não
31	OBITOPUERP	C(01)	Morte durante o puerpério, conforme tabela- 9- Ignorado 1- Sim, ate 42 dias 2- Sim, de 43 dias
32	ASSISTMED	C(01)	Indica se houve assistência medica, conforme a tabela- 9- Ignorado 1- Com assistência 2- Sem assistência
33	EXAME	C(01)	Indica se houve exame complementar, conforme a tabela- 9- Ignorado 1- Sim 2- Não
34	CIRURGIA	C(01)	Indica se houve cirurgia, conforme a tabela- 9- Ignorado 1- Sim 2- Não

35	NECROPSIA	C(01)	Indica se houve necropsia, conforme a tabela- 9- Ignorado 1- Sim 2- Não
36	LINHAA	C(20)	Linha A do atestado, conforme a Classificação Internacional de Doença (CID), 10a. Revisão.
37	LINHAB	C(20)	Linha B do atestado, conforme a Classificação Internacional de Doença (CID), 10a. Revisão.
38	LINHAC	C(20)	Linha C do atestado, conforme a Classificação Internacional de Doença (CID), 10a. Revisão.
39	LINHAD	C(20)	Linha D do atestado, conforme a Classificação Internacional de Doença (CID), 10a. Revisão.
40	LINHA	II	C(20) Linha II do atestado, conforme a Classificação Internacional de Doença (CID), 10a. Revisão.
41	CAUSABAS	C(04)	Causa básica, conforme a Classificação Internacional de Doença (CID), 10a. Revisão
42	TPASSINA	C(01)	-
43	DTATESTADO	C(08)	Data do Atestado .
44	CIRCOBITO	C(01)	Indica o tipo de acidente, se cabível- 9- Ignorado 1- Acidente 2- Suicídio 3- Homicídio 4- Outros
45	ACIDTRAB	C(01)	Indica se foi acidente de trabalho, conforme a tabela- 9- Ignorado 1- Sim 2- Não
46	FONTE	C(01)	Fonte da informação, conforme a tabela- 9- Ignorado 1- Boletim de Ocorrência 2- Hospital 3- Família 4- Outra
47	TPPOS	C(01)	Óbito investigado 1-Sim, 2- Não
48	DTINVESTIG	C(08)	Data de investigação.
49	CAUSABAS_O	C(04)	Causa básica original, a primeira informação que entra no sistema.
50	DTCADASTRO	C(08)	Data de cadastro do registro no sistema.
51	ATESTANTE	C(01)	Indica se o medico que assina atendeu o paciente 1- Sim 2- Substituto 3- IML 4- SVO 5-
52	FONTEINV	C(01)	Fonte de investigação 1 Comitê de Morte Materna e/ou Infantil 2 Visita domiciliar / Entrevista família 3 Estab
53	DTRECEBIM	C(08)	Data de recebimento no nível central, data da última atualização do registro.
54	UFINFORM	C(02)	Código da UF que informou o registro.
55	CODINST	C(14)	Código da instalação da geração dos registros.
56	CB_PRE	C(04)	Causa selecionada sem re-seleção (novo SCB).

Considerando que a mortalidade e a letalidade do câncer são parâmetros importantes para a avaliação da gravidade da endemia, do retardo na

detecção de casos, do início do tratamento e da sua efetividade, a utilização do SIM é de extrema relevância para subsidiar o sistema de vigilância hoje existente (MINISTÉRIO DA SAÚDE - Manual_Operacional_APAC_v_1_1.pdf [Internet], n.d.; MINISTÉRIO DA SAÚDE - Manual_Operacional_BPA.pdf [Internet], n.d.; MINISTÉRIO DA SAÚDE - Manual_Operacional_SIA_v_1.pdf [Internet], n.d.; MINISTÉRIO DA SAÚDE, SVS, Brasil, n.d.).

8. Pareamento de Registros

O pareamento de registros, ou *record linkage*, como foi denominado por Winkler 1969 (Fellegi and Sunter, 1969), é a tarefa de vincular, de alguma maneira, informação de duas ou mais fontes de dados que representem uma mesma entidade. O pareamento de registos também é conhecido como vinculação de dados, ou ainda em inglês, *data linkage*, *record matching*, *database matching*, *data cleaning*, *data scrubbing*, *data standardization*, ETL (*extraction, transformation and loading*) ou *merge/purge problem*.

Do ponto de vista funcional, o pareamento de registros é utilizado para a deduplicação de um conjunto de dados (*data set deduplication*) e a vinculação de dois bancos de dados (*data set linkage*) – formando um conjunto de dados principal com informações dos bancos de origem.

Hoje podemos observar que o pareamento de registros se encontra difundido nas seguintes áreas: epidemiologia (orientada para pacientes e estudos longitudinais), estatísticas do censo, listas de contatos comerciais (limpeza e atualização), detecção de crimes e fraudes. Em estudos epidemiológicos, o pareamento de registros é frequentemente utilizado como o passo inicial para a análise de dados, organização das informações em série histórica (Moura, Prestes, Duncan, and Schmidt, 2014) e/ou projetos de mineração de dados (Torra and Domingo-Ferrer, 2003; Winkler, 2006).

Os tipos fundamentais de pareamento de registros são dois: o pareamento de registros determinístico e o pareamento de registros probabilístico. Em sua versão determinística, o pareamento de registro vinculará os registros de duas bases de dados apenas se o valor existente nessas duas variáveis for exatamente igual. Por exemplo, consideremos que desejamos vincular duas bases de dados, e a busca do

vínculo se dá pelo nome do paciente. Um par de registro só será vinculado se for exatamente igual. O pareamento de registros probabilístico, por outro lado, utiliza um conjunto de variáveis do conjunto de dados, convertendo similaridade das diversas informações em uma probabilidade. A partir dessa probabilidade é que um par de registro poderá ser classificado como uma ligação (ou seja, pertencem a uma mesma entidade), ou como uma não ligação (Blakely and Salmond, 2002).

8.1 Etapas do pareamento de registros

O problema que se coloca muitas vezes ao tentarmos realizar um pareamento de registros é que as informações sobre uma entidade podem se encontrar em mais que uma fonte de dados. Assumindo que a fonte de dados apresenta um identificador único livre de erro para a formação dos pares, um simples relacionamento determinístico seria suficiente. Em verdade, temos que no mundo real as fontes de dados são "sujas", isto é, apresentam distorções na informação armazenada – isso pode se dever a erros de digitação, troca de registros, ausência da informação por má digitação, não uniformidade do formato do dado, abreviações, mudanças sobre o tempo, etc.

O pareamento de registro pode ser considerado como uma parte de um processo de limpeza dos dados, um passo fundamental no processo para se obter conhecimento(Gu and Baxter, 2006). Independentemente da técnica utilizada para o pareamento dos registros, um fluxo de trabalho (workflow) foi estabelecido, trabalho de vários autores (Fellegi and Sunter, 1969; Winkler, 1999; Baxter, Christen, and Churches, 2003; Furht and Escalante, 2011), como boa prática para se obter resultados mais consistentes. Esse fluxo de trabalho é bem apresentado pela implementação e esquema dado por (Elfeky, Verykios, and Elmagarmid, 2002; Baxter, Christen, and Churches, 2003; Christen, Churches, and Hegland, 2004; Christen, 2008*a*).

O processo do pareamento de dados (FIGURA 1) se inicia com a aquisição das bases de dados. Os seus registros passam então por uma avaliação da qualidade da informação e também padronização das mesmas. Para reduzir o número de comparações, são implementadas técnicas de bloqueamento e busca para a formação dos pares de registros. Os pares de registros são então convertidos em vetores de comparação, sobre os quais as funções de comparação são aplicadas e o resultado

é passado para o modelo de decisão, que determina o estado final do par de registro (ligação verdadeira, ligação falsa, ou provável ligação). A última etapa do processo consiste em avaliar a qualidade dos resultados gerados(Handbook of data quality: research and practice, 2013).

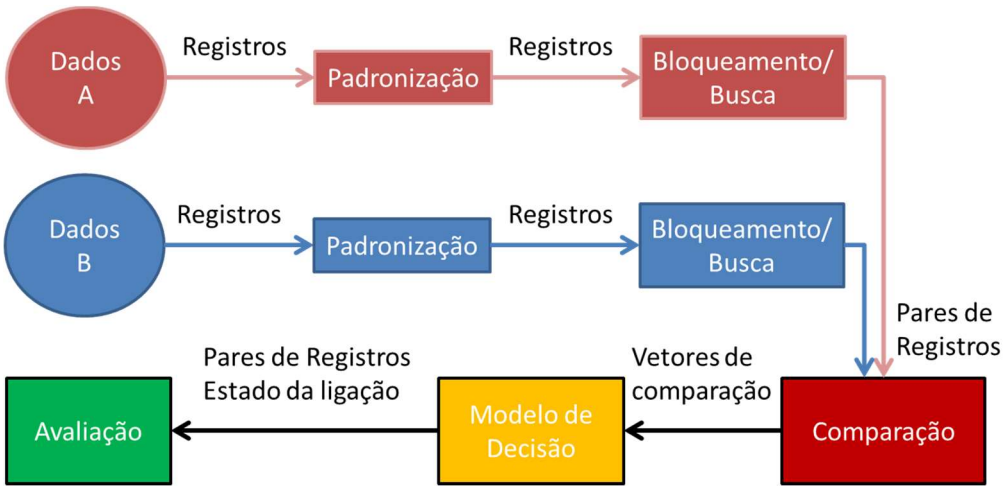


FIGURA 1 – Diagrama do fluxo da informação de um sistema de pareamento de registros.

8.2 Padronização

As bases de dados podem conter em suas variáveis valores ausentes (*missing values*), erros e outras informações expressas em mais que um formato – uma data de nascimento pode ser apresentada no formato DDMMYYY ou MM/DD/YYYY. Assim a padronização (*Standardization*) é um passo da limpeza de dados, uma boa prática para qualquer pareamento de registros. A padronização tem grande impacto na qualidade do pareamento de registros, pois evita que seja detectada uma dissemelhança entre as informações comparadas apenas por uma variação de formato.

Entre as principais padronizações executadas no pareamento de registros temos a eliminação de erros ou variações tipográficas (através da verificação de frequência dos valores, substituir “Mraia” por “Maria”); padronização de uma sequência de texto (*string*) em caixa alta e sem caracteres especiais – como acentos, cedilha –; conversão de variáveis numéricas e de datas para um único formato entre fontes de dados; expansão de abreviações seguida de sua substituição por uma forma padronizada; substituição de apelido pelo nome em sua forma regular (Bob por Robert,

Zé por José, Tião por Sebastião) (Herzog, Scheuren, and Winkler, 2007; Handbook of data quality: research and practice, 2013).

Em geral, para problemas envolvendo pareamento de registros de pacientes, executa-se a padronização das bases com a transformação da data de nascimento em um campo caractere com seis ou oito posições; com a utilização uma mesma regra de codificação do campo sexo nos diferentes arquivos; com a transformação de todos os caracteres alfabéticos da forma minúscula para a maiúscula, seguida da eliminação de caracteres de pontuação e a eliminação de espaços em branco no início do campo.

Os procedimentos para a padronização também podem contar com a substituição de caracteres especiais (âàãá ÁÂÀÃéêÉÊíÍóôõÓÔÕüúÜÚ Ççñ !@#$%&*()-=_+¹²³£¢¬\|/?;:. >,<]}~^[{´`ºªª§) por traço ("-" ou "*"), a utilização de caixa alta para todas as letras, a eliminação de caracteres de pontuação, acentuação e espaços em branco, bem como a exclusão de todas as preposições "de", "do", "da", "dos", "das" e similares (Stevens, Schmidt, and Duncan, 2014).

8.3 Bloqueamento

O bloqueamento é uma técnica de organização dos dados que, ao dividir a base de dados, segundo algum critério, em menores blocos de registros., reduz o espaço de comparações entre pares de registros. Ela tem por finalidade reduzir o número de comparações no pareamento de dados.

A comparação natural dos registros de duas bases de dados, ou seja, uma comparação de todos contra todos, é uma tarefa que implica o crescimento em valor quadrático do número de pares de registros a serem comparados. Sendo assim, para grandes bases de dados, o pareamento de registros se torna não factível computacionalmente: pode ser computacionalmente muito custoso (uma tarefa que não pode ser realizada pelo parque computacional disponível), ou necessitará de um tempo para execução além do disponível para apresentação dos resultados.

Se por um lado o bloqueamento pode gerar uma redução do número total de comparações a serem feitas, por outro isso implica que registros de uma mesma entidade poderão não virem a ser comparados. Um exemplo para essa situação, consideremos um bloqueamento pela primeira letra do nome de paciente. Apenas serão

comparados os registros iniciados com a mesma primeira letra. Se existirem dois registros da mesma paciente, sendo um “HELENA DA SILVA” e outro “ELENA DA SILVA”, pelo bloqueamento realizado, eles não seriam comparados. Situação que pode, em alguns casos, ser evitada com padronização de nomes e variantes, ou aplicação de um algoritmo para tratamento de fonemas – como o Soundex da Caixa Econômica Federal para nomes brasileiros(Freire et al., 2009).

Além disso, ao se definir as chaves de um bloqueamento, existe uma relação de custo-benefício a ser considerado: se os blocos resultantes contiverem um grande número de registros, então mais pares de registros necessariamente serão formados. Isso leva a um grande número de comparações, portanto ineficiente. Por exemplo, se gênero (masculino, feminino) for usado como atributo de bloqueamento, os registros serão alocados em apenas dois grandes blocos. Por outro lado, se os blocos de registros forem muito pequenos, então verdadeiros pares de registros poderão não ser comparados, reduzindo assim a acurácia (sensibilidade) do pareamento de registros(Baxter, Christen, and Churches, 2003).

O bloqueamento consiste essencialmente em agrupar os registros em blocos/conglomerados segundo uma variável da base de dados. Assim a totalidade dos registros formam blocos, uma partição do conjunto de dados, e apenas os registros de mesmo bloco são comparados. Um exemplo pode ser visto na FIGURA 2.

Podemos observar que o bloqueamento pelos 3 primeiros dígitos do CEP realiza uma redução de 30 para 11 pares para serem comparados. Esse tipo de bloqueamento é conhecido como bloqueamento padrão. Outra aplicação muito comum de bloqueamento é o uso das primeiras letras de nomes, ou endereços(Baxter, Christen, and Churches, 2003). A aplicação de Soundex – um algoritmo fonético para indexação de nomes por som – é também um método eficiente para formação de blocos e permite que se fuja da perda de qualidade do pareamento decorrente de erros de digitação ou pronúncia (Alvey, 1997; Randall, Ferrante, Boyd, and Semmens, 2013).

FIGURA 2 – Exemplo da redução do número de comparações por bloqueamento dos 3 primeiros dígitos do CEP.

Outro tipo de bloqueamento é o método de *sorted neighborhood*, que pode ser aplicado como método de busca também. Esse método ordena os registros tendo por base uma chave para o ordenamento e então movimenta uma janela – um conjunto de registros – de tamanho fixo (*w*) sequencialmente sobre os registros ordenados. Os registros da janela são então pareados com cada um dos outros e armazenado o par candidato em uma lista. O uso da janela limita o número de possíveis comparações de pares de registros – assim, para cada registro da janela existirão *2w-1* comparações a serem feitas. Se considerarmos o pareamento de duas bases de dados com *n* registros em cada, por este método, teremos um total de *wn* comparações (Christen, 2008*b*). A contrapartida da redução do número de comparações com o uso do método *sorted neighbourhood* surge quando um número de registros maior que o

tamanho da janela tem o mesmo valor em uma chave de ordenamento. Por exemplo, consideremos o ordenamento por sobrenome, centena de registros podem ter o valor "SILVA", e se o tamanho da janela é menor, nem todos os registros com o sobrenome "SILVA" serão comparados(Baxter, Christen, and Churches, 2003).

8.4 Metodologia de busca para formação de pares

O método de busca para formação de pares de registros é uma técnica de organização dos dados que tem por finalidade reduzir o número de comparações do pareamento. Esses métodos consistem em algoritmos para a determinação de que pares das fontes de dados A e B devem ser comparados, pela determinação de que atributos serão usados para organizar as comparações.

```
For each tuple a in A do
    For each tuple b in B do
      If a and b satisfy the join condition
        Then output the tuple <a,b>
```

QUADRO 1 – Algoritmo nested loop join (NLJ).

Entre os métodos de busca mais utilizados encontramos *o nested loop join* (NLJ) e o *sorted neighborhood method* (SNM). O NLJ consiste na formação de pares a serem comparados a partir de todos registros de uma base A comparados com todos os outros da outra base B . Os pares de registros, que podem ser representados por uma tupla, são gerados pelo algoritmo como ilustrado em QUADRO1.

Quando lidamos com pequenas bases de dados a terem seus registros pareados, o método NLJ pode ser aceitável. Entretanto, para grandes bases de dados ele se torna deveras oneroso do ponto de vista computacional e de tempo – lembremo-nos sempre que o crescimento do número de pares é exponencial. Sendo assim, o *sorted neighborhood method* (SNM) é muito mais adequado, pois promove uma redução no número de pares formados a serem comparados posteriormente.

As etapas executadas no SNM (FIGURA 3), consistem de ordenar os registros da base de dados A e B segundo os atributos selecionados. A isso segue

a determinação do tamanho das janelas de registros para comparação - ωA e ωB. Assim a formação de pares de registros se dá ao fazer avançar essas janelas sobre as bases de dados. Pelo fato do ordenamento das bases de dados, segundos os critérios escolhidos, colocarem próximos registros parecidos, espera-se que serão comparados apenas registros que estiverem próximos, dentro de ωA e ωB . Pela determinação dos tamanhos de ωA e ωB será ajustada a quantidade de pares que serão formados, isso implica reduzir o número de pares para comparação(Christen, 2008*b*; FRIL: Fine-grained record linkage software [Internet], 2011) – o que é considerado um benefício -, mas ao custo de que, se mal ajustados os tamanhos das janelas, alguns pares venham a jamais serem comparados.

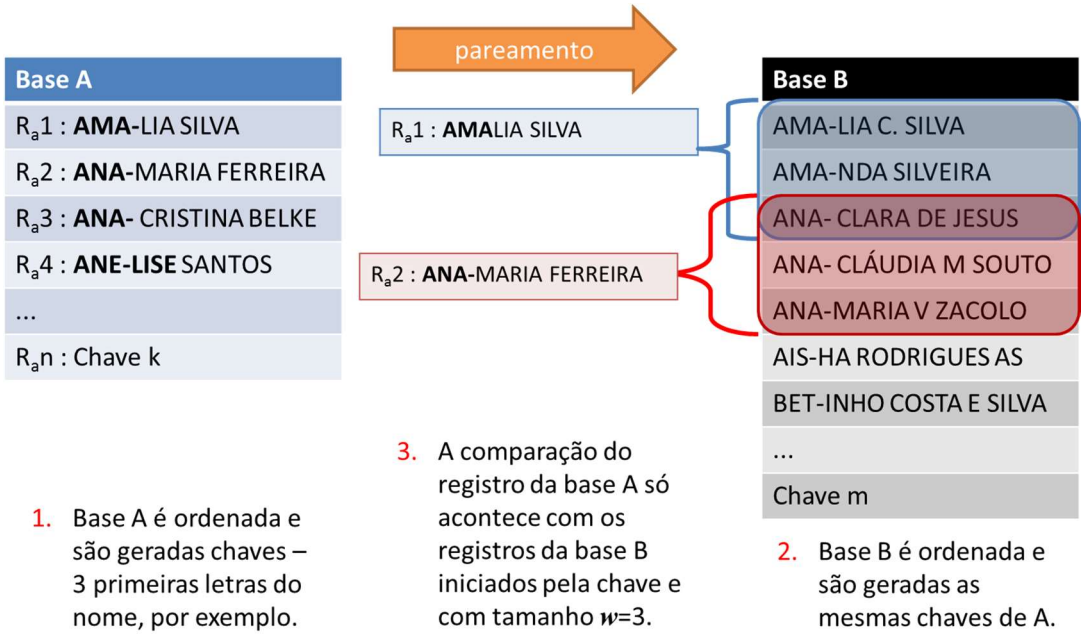


FIGURA 3 – Dinâmica da busca dos registros para formação de pares pelo método SNM.

8.5 Metodologia de comparação de registros

Os pares de registros formados pela metodologia de bloqueamento e busca dão origem aos vetores de comparação. Sobre esses vetores serão aplicadas as funções de comparação – os resultados dessas funções serão então passados ao modelo de decisão para que possa ser realizada a classificação do par de registros, ou seja, a sua classificação final em ligação (registros de uma mesma entidade), possível ligação ou não ligação.

As medidas usadas para comparação das sequências de caracteres no pareamento de registros podem ser alocadas em duas classes: as funções de distância e medidas de similaridade baseada em símbolos – segmentos de uma sequência de caracteres. As funções de distância mapeiam um par de *strings* em um número real, onde, em geral, um pequeno valor significa grande similaridade e um valor grande significa pouca similaridade. Assim, o inverso de uma função de distância pode se converter em função de similaridade – neste caso quanto maior o valor, maior a similaridade (Cohen, Ravikumar, and Fienberg, 2003). As medidas de comparação que tem por base segmentos de uma sequência de caracteres, consideram que uma *string* pode ser decomposta em subconjuntos de caracteres, ou palavras e mapeiam esses subconjuntos no espaço dos números reais. Ou seja, avaliando a similaridade entre os subconjuntos formados para as duas *strings*, um número real é atribuído a essa maior semelhança ou dissimilaridade (Cohen, Ravikumar, and Fienberg, 2003).

8.5.1 Distância de Levenshtein

A distância de Levenshtein é uma medida de similaridade para uma sequência de caracteres texto, comumente chamada *string*. Explicada de outra maneira, essa distância é uma métrica para mensurar a diferença entre duas sequências de caracteres – altamente relacionada com métodos de alinhamento pareado de sequências, os quais são usados para encontrar a melhor correspondência por partes em uma sequência de caracteres. A distância de Levenshtein busca encontrar a edição mínima que uma string deve sofrer para se tornar igual à string de referência. Para isso serão consideradas inserções de caracteres, exclusão de caracteres e substituições até que uma *string* se transforme na outra, passem a ser idênticas. Essa medida de similaridade foi desenvolvida por Vladimir Levenshtein em 1965 (Levenshtein, 1966) – precedendo assim o trabalho da formalização do pareamento de registros datado de 1969 (Fellegi and Sunter, 1969). A distância de Levenshtein é também conhecida como distância de edição, porém é necessário ter cuidado com essa nomenclatura, pois distância de edição é uma classe de métricas de distâncias, entre as quais se encontra a distância de Levenshtein (Ristad and Yianilos, 1998).

A distância de edição de Levenshtein é definida matematicamente entre duas strings a, b como $lev_{a,b}(|a|,|b|)$ onde

$$lev_{a,b}(i,j) = \begin{cases} \max(i,j), \text{ se } \min(i,j) = 0 \\ \min \begin{cases} lev_{a,b}(i-1,j)+1 \\ lev_{a,b}(i,j-1)+1 \\ lev_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} \text{, caso contrário.} \end{cases} \quad (0.1)$$

onde $1_{(a_i \neq b_j)}$ é uma função indicadora igual a zero quando $a_i = b_j$ e igual a 1 em caso contrário. É interessante observar que o primeiro elemento no mínimo da função corresponde a remoção de um caractere de a em relação a b, o segundo a inserção de um caractere e o terceiro uma mudança de posição de um caractere – dependendo de onde o respectivo caractere se encontra, ou se são os mesmos.

		k	**i**	**t**	**t**	**e**	**n**
	0	1	2	3	4	5	6
s	1	1	2	3	4	5	6
i	2	2	1	2	3	4	5
t	3	3	2	1	2	3	4
t	4	4	3	2	1	2	3
i	5	5	4	3	2	2	3
n	6	6	5	4	3	3	2
g	7	7	6	5	4	4	3

		S	**a**	**t**	**u**	**r**	**d**	**a**	**y**
	0	1	2	3	4	5	6	7	8
S	1	0	1	2	3	4	5	6	7
u	2	1	1	2	2	3	4	5	6
n	3	2	2	2	3	3	4	5	6
d	4	3	3	3	3	4	3	4	5
a	5	4	3	4	4	4	4	3	4
y	6	5	4	4	5	5	5	4	3

QUADRO 2 – Distância de Levenshtein computada em matrizes.

Segue o exemplo clássico dado para a distancia de Levenshtein:

kitten → sitten (substituição de "s" por "k")

sitten → sittin (substituição de "i" por "e")

sittin → sitting (inserção de "g" no final da sequência).

A distância de Levenshtein pode ser vetorizada, resolvidas a partir de matrizes. O conteúdo dessas matrizes mostra o escore dessa distância para cada caractere da sequência a em relação os caracteres da sequência b, como mostra o exemplo acima – QUADRO 2.

Essa abordagem por matrizes é, do ponto de vista computacional, muito mais interessante, pois permite um ganho de velocidade proporcionado por sistemas vetoriais(Wagner and Fischer, 1974). O princípio de invariância é mantido pelo algoritmo, sempre usando o mínimo da linha ou da coluna para a soma final do escore de similaridade(Cohen, Ravikumar, and Fienberg, 2003).

8.5.2 Distâncias numéricas

As distâncias numéricas podem ser aplicadas para medidas pelo menos discretas e ordinais. Elas funcionam a partir do cálculo da diferença entre dois valores observados na mesma métrica(Cohen, Ravikumar, and Fienberg, 2003). Por exemplo, ao comparar duas datas no formato YYYY-MM-DD, ou em formato data-time em Linguagem de Consulta Estruturada (*Structured Query Language* – SQL), a diferença entre elas é apresentada em dias.

8.5.3 Distância de Jaro-Winkler

A distância de Jaro-Winkler (Winkler, 1990) é uma medida de similaridade entre duas sequências de caracteres, uma variação da medida de edição de sequência de caracteres desenvolvida por Jaro. Elas foram desenvolvidas para serem aplicadas em pareamento de registros, sendo melhor aplicada em situações de *strings* curtas – tal como nome de pessoas decompostos em nome, nome do meio e sobrenome. Essas medidas são amplamente utilizadas no pareamento de registros até os dias de hoje. Quanto maior for a medida de Jaro-Winkler, maior será a similaridade entre as *strings* comparadas. O escore para essa distância vai de 0 (sem qualquer similaridade entre *strings*) a 1 (similaridade máxima, identidade)(Winkler, 1990).

Matematicamente, a distância de Jaro-Winkler d_j para duas sequência de caracteres a e b é

$$d_j = \begin{cases} 0, \text{ se } m = 0; \\ \frac{1}{3}\left(\frac{m}{|a|} + \frac{m}{|b|} + \frac{m-t}{m}\right) \text{ em caso contrário.} \end{cases} \quad (0.2)$$

onde m é o número de caracteres idênticos entre as *strings* e t é a metade do número de transposições na sequência. Dois caracteres de a e b são considerados idênticos apenas se eles são os mesmos e não mais que

$$\left|\frac{\max(|a|,|b|)}{2}\right| - 1 \quad (0.3)$$

Em resumo, cada caractere de a é comparado com todos caracteres correspondentes em b ; então o número de semelhanças de caracteres (mas em ordem diferentes) é dividida por 2 e define o número de transposições. Por exemplo, ao se comparar CRATE com TRACE, apenas os caracteres R,A e E são correspondências exatas – logo $m = 3$. Os caracteres C e T estão presentes em ambas *strings* e não vão além de 1 aparição. Disso decorre que a parte inteira de 5/2-1=1, fazendo com que $t = 0$. Em IsAIAS e IzAIhAS as letras correspondentes são IAIAS e se encontram na mesma sequência, logo nenhuma transposição é necessária.

A distância de Jaro-Winkler utiliza um parâmetro escalar de prefixo p, o que possibilita melhores classificações para *strings* que apresentam concordância no início de sua sequência de caracteres de tamanho l . Consideremos duas *strings* a e b , sua distancia de Jaro-Winkler é dada por d_w . Tal que

$$d_w = d_j + \left(lp\left(1 - d_j\right)\right) \quad (0.4)$$

onde d_j é a distância de Jaro-Winkler para as *strings* a e b ; l o comprimento do prefixo comum da *string* até um máximo de quatro caracteres; p é um fator escalar constante para quando do escore será ajustado pela existência de um fator comum entre as *strings* avaliadas.

O valor de p nunca deve ultrapassar o valor de 0,25, pois nestes casos a distância pode apresentar uma valor maior que 1 – perdendo assim a sua

normalização. O valor padrão utilizado para a distancia de Jaro-Winkler é $p = 0,1$ (Cohen, Ravikumar, and Fienberg, 2003).

8.5.4 n-GRAM

Um n-gram é o conjunto de todas as *substrings* (partição de uma sequencia de caracteres) que podem ser geradas a partir de uma determinada *string* e "q" representa o tamanho destas *substrings*. O algoritmo de n-gram é utilizado como uma técnica de filtragem, cuja finalidade é utilizar para comparação apenas subconjuntos onde pode teoricamente existir ligação entre palavras (matching)(Kondrak, 2005) – no caso do pareamento de registros blocos(Practical Cryptography [Internet], n.d.)

Entretanto, um modelo n-gram, é um tipo de modelo probabilístico de linguagem que funciona como um modelo de Markov de ordem (n-1), podendo apontar portanto a probabilidade de ocorrência de uma sequência de caracteres em função de outra – funcionando assim como uma medida de similaridade normalizada, ou seja, que assume valores entre 0 e 1(Jurafsky and Christopher Manning, n.d.). Em resumo, a técnica pode ser aplicada de forma a identificar sub-sequências de caracteres que possuam elementos em comum.

Sejam *strings* A e B podemos gerar os n-grams de A e B. A seguir, contamos o número de n-grams idênticos. Assim é possível encontrar a distância "n-gram":

$$|\text{distância n-gram}| = |\text{total de n-gram formados}| - |\text{número de n-gram coincidentes}| \quad (0.5)$$

Exemplo de bi-gram para a palavra "Isaias" e "Izaias":

{"#i, is, sa, ai, ia, as, s#" }

{"#i, iz, za, ai, ia, as, s#"}

Realizando os cálculos temos que

| distância bi-gram| = 7 - 5 = 2.

Portanto uma dissimilaridade baixa (2/7 = 0,28571), ou uma alta similaridade (1 - 0,28571 = 0,71428) posto que as palavras são muito parecidas.

Sejam $|x|$ e $|y|$ os tamanhos das *strings* x e y, que serão comparadas. O tempo decorrido para execução de um n-gram é melhor do que o da programação dinâmica, pois como cada n-gram de x é comparado apenas com uma certa quantidade dos n-gram de y então o tempo gasto é da ordem de $O\left(\min\left(|x|,|y|\right)\right)$ e a quantidade de espaço requerida é da ordem de $O\left(|x|+|y|\right)$. Este algoritmo é mais eficiente do que o de programação dinâmica, mas a qualidade do escore gerado não é tão específico, sendo altamente sensível(Foster Jr and Evans, 2003).

8.5.5 Outras distâncias e suas propostas aplicações

A *Soundex Distance Metric* é uma técnica que reduz cada sequência de caracteres a um código soundex (*soundex code*). O código soundex é formado por uma letra e três dígitos, de forma que considera similar todas as *strings* que possuem o mesmo código(Hall and Dowling, 1980). A implementação do algoritmo soundex tem um bom desempenho, uma vez que é executado com tempo linear. Recomenda-se que o soundex só deve ser aplicado para verificar erros fonéticos, porque grandes erros ocorrem quando usado para outra aplicação qualquer – como codificação de uma longa *string* com várias palavras. Embora o soundex tenha sido implementado para dar conta da fonética da língua inglesa, para ser utilizado na Língua Portuguesa, é possível utilizar uma adaptação criada pela Caixa Econômica Federal(Freire et al., 2009).

De forma detalhada, temos que a distância soundex tem por finalidade transformar um nome em um código de 4 dígitos, de forma tal que sons similares possuam estes 4 caracteres. O primeiro caractere é a primeira letra do nome, as letras seguintes são substituídas pelos seguintes números:

0) A, E, I, O, U, H, W, Y
1) B, P, F, V
2) C, S, K, G, J, Q, X, Z
3) D,T
4) L
5) M, N
6) R

Em seguida os zeros são removidos; as repetições do mesmo dígito são reduzidas a um único dígito; e, por fim, o código é truncado a apenas uma letra e 3 dígitos.

O exemplo clássico de apresentação da distância métrica soundex apresenta bem a técnica: “Dickson” e “Dixon” recebem o mesmo código D25, o que mostra a sua similaridade.

Outra técnica para comparação de sequências de texto, embora pouco aplicada e reconhecida como métrica para avaliação, é o Bit-Parallelism. O Bit-Parallelism procura explorar o paralelismo dos computadores quando estes trabalham com bits. Os dois algoritmos mais comumente implementados são os automatos não determinísticos e programação dinâmica de matrizes (Levenshtein). Um ambiente de grande aplicação se dá em sistemas que suportem execução paralela, ambientes multiprocessados. Formalmente, seja w o tamanho de uma *computer word* (a unidade natural de dado usada por uma arquitetura de computador particular) que pode ser de 32 ou 64 bits. Uma máscara pode ser usada para representar diversas palavras utilizando uma única *computer word* isto possibilita que sejam feitas ligações (matches) em conjuntos de caracteres em vez de ser feito em um único caractere. Esta técnica melhora a eficiência dos algoritmos de Levenshtein e NFA, e os valores dos escores obtidos nas comparações não são modificados (Herzog, Scheuren, and Winkler, 2007).

8.6 Preservação da privacidade das informações em pareamento de registros

Amplamente utilizado para diversas aplicações, o pareamento de registros, por exemplo, quando usado em base de dados médicas, pode envolver informações pessoais sensíveis sobre o paciente. Por isso ganha importância o problema do pareamento de registros que zelem pela manutenção da privacidade. As noções de privacidade da informação hoje, fundamentalmente no que concerne ao problema do pareamento de registros privado, não tem uma única interpretação, tampouco abordagem na comunidade científica(Bonomi, Xiong, and Lu, 2013). A busca pela privacidade do pareamento de registro tem sido feita no sentido de criar modelos de protocolos, tipos de objetos a serem usados no pareamento de registros. Uma forma de classificar distintamente essas técnicas é a seguinte: transformações

seguras, *secure multiparty computation* e métodos híbridos(Bonomi, Xiong, Chen, and Fung, 2012).

A transformação segura consiste em realizar o pareamento de registros após os dados sofrerem alguma espécie de transformação. Nesta modalidade os dados são codificados na origem dos dados, usando uma técnica de transformação segura dos dados, para então serem enviados a um local para a realização do pareamento – cuja única finalidade é realizar o pareamento da base. Neste enfoque as duas principais estratégias propostas são *hashing* e *embedding*. Ambas soluções se baseiam em funções *hash* que, após codificação, tentam ligar os dados entre as duas bases pareadas, computando assim a sua medida de similaridade. Um problema desse enfoque decorre do fato que essas funções não oferecem limites formais para as medidas de similaridade dentro do espaço *hash* onde operam. Transformada a informação, não se pode garantir que a transformação preserve as características próprias de similaridade para a informação codificada. Além disso, esse enfoque está sujeito a um *dictionary attack* (Bonomi, Xiong, Chen, and Fung, 2012)– uma técnica para violar um mecanismo de codificação ou autenticação pela tentativa de determinar a sua chave de decodificação ou a frase secreta, tentando centenas ou às vezes milhões de possibilidades prováveis, como palavras em um dicionário(Adams, 2011). Uma vez violado conjunto de funções *hash*, a segurança do sistema e privacidade estão arruinadas. As principais técnicas usando funções *hash* são *Bloom filters*, *q-grams hashing* e TFIDF *hashing*(Schnell, Bachteler, and Reiher, 2009)(Bonomi, Xiong, Chen, and Fung, 2012)(Practical Cryptography [Internet], n.d.). A técnica que usa a classe de funções embedding, é também uma codificação, mas que tem a vantagem de que as distâncias computadas sobre o espaço de caracteres original podem ser colocadas em relação com a distância no novo espaço após transformação. A qualidade dessa relação se dá apenas em função da distorção introduzida pelo mapeamento *embedding* utilizado. Entretanto, funções *embedding* são computacionalmente muito dispendiosas. Outra desvantagem é que dependendo da distorção introduzida, o problema de codificação se torna um grande desafio técnico.

A *Secure Multiparty Computation* é uma técnica que converte o problema do pareamento de registros em um ambiente de comunicação segura. O pareamento de registro tem suas fontes de informação adaptadas para um protocolo de

comunicação usando criptografia – de forma que só são repassadas a outras fontes os dados estritamente necessários para que o pareamento seja realizado. O custo computacional e de comunicação dessa técnica é consideravelmente alto em aplicações reais.

Os métodos híbridos combinam o processo de anonimização ou transformações seguras com técnicas de *Secure Multiparty Computation*. Dessa forma o custo – principal problema dessa técnica – é reduzido, tornando relativamente viável a aplicação da combinação de métodos (Bonomi, Xiong, Chen, and Fung, 2012).

8.6.1 Filtros de Bloom (Bloom Filters)

A vinculação de bases de dados por meio do pareamento de registros, usando informações particulares de pessoas – identificação, padrões de comportamento, financeiras, etc – tem tido um crescente interesse científico e comercial nos últimos anos. Em algumas aplicações, a vinculação de bases de dados tem sido realizada sem o uso de um único identificador, um único número pessoal. Para contemplar itens de segurança para manutenção de dados pessoais, surgiu o que chamamos pareamento de registros preservando privacidade - *privacy preserving record linkage* (PPRL). Dados de variáveis chamadas quasi-identificadores, ou seja, características pessoas que possibilitam a identificação de uma pessoa, como primeiro nome, sobrenomes, data de nascimento, têm sido amplamente aplicadas para vincular bases de dados distintas sem qualquer identificador pessoal único. O problema existente nos quasi-identificadores é que eles contém erros e, para alguns problemas de relacionamento erros de vinculação são inaceitáveis(Schnell, Bachteler, and Reiher, 2009).

Uma primeira proposta para a realização de pareamento de registros preservando dados particulares apareceu em 2009, com a aplicação de *Bloom filters*. Neste protocolo, curadores da fonte de dados A e B inicialmente concordam adotando uma senha para comunicação. A seguir, eles padronizam seus identificadores, adicionando espaços em branco no início e no final da informação; então as separam em *substrings* de dois caracteres (bi-grams). Os bi-gram de um identificador são mapeados por funções hash para um *bloom filter* – segundo códigos chaveados de autentificação de mensagem (HMACs), muito parecida com chaves MD5 ou SHA-1.

É computada então a similaridade entre dois *Bloom filters* pela avaliação do número posições de bits coincidentes, que são ajustados para 1. A similaridade dos identificadores codificados pode ser feita, de forma aproximada, pela métrica de similaridade do índice de Jaccard, ou coeficiente de Dice ou Tanimoto (Schnell, Bachteler, and Reiher, 2009). No final, esse procedimento possibilita a realização do pareamento de registros criptografados mesmo com erros existentes no conteúdo das informações dos pares. Qualquer um que leia os dados transformados em um *Bloom filter*, verá apenas zeros e uns, sendo necessário um grande trabalho computacional para conseguir reverter esses zeros e uns na informação original(Schnell, Bachteler, and Reiher, 2009).

Um *Bloom filter* de um nome é um conjunto de partições atômicas da sequência completa desse nome. Ou seja, elementos derivados da formação de bigrams desse nome. Consideremos o nome MULLER. Temos, portanto, os bigrams

{#M, MU, UE, EL, LL, LE, ER, R#}.

Assim o bigram ER pode ser mapeado com k=15 funcoes hash formando o vetor exibido na FIGURA 4 abaixo:

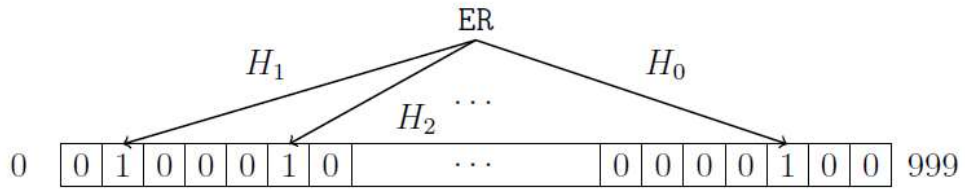


FIGURA 4 – Filtro de Bloom do bigram ER.
FONTE (Cryptanalysis of basic Bloom Filters used for Privacy Preserving Record Linkage - -download=wp-grlc-2014-04.pdf [Internet], n.d.).

Ao se mapear cada um dos bigram com 15 funções *hash*, temos resultados dos 8 átomos. Eles são então combinados com operador *bitwise* OR (ou),

que resulta no Bloom filter para o nome MULLER (FIGURA 5):

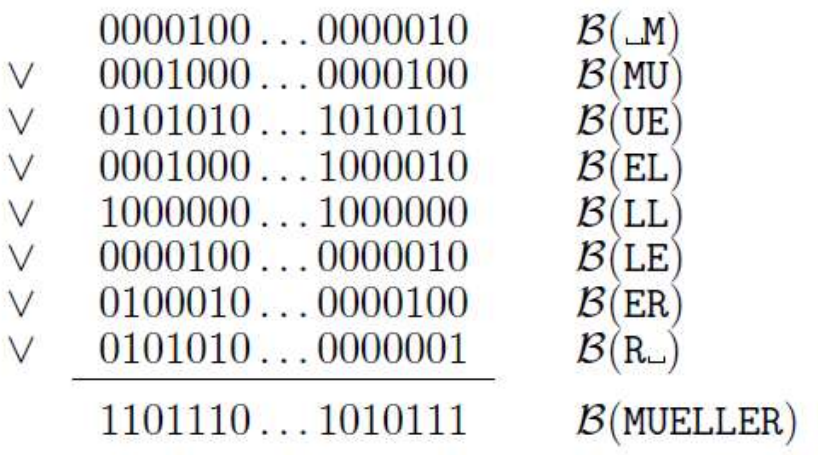


FIGURA 5 – Bloom filter da palavra MULLER, formado pela codificação de todos seus bigram. FONTE (Cryptanalysis of basic Bloom Filters used for Privacy Preserving Record Linkage - -download=wp-grlc-2014-04.pdf [Internet], n.d.).

8.7 Modelo de Decisão

Antes de podermos falar de modelos de decisão – ou seja, um modelo que nos orientará em relação a como classificar um par de registros – precisamos definir um conjunto de estruturas. Consideremos duas fontes de dados A e B . O conjunto ordenado dos pares de registro $A \times B = \{(a,b) : a \in A, b \in B\}$ é a união de dois conjuntos disjuntos, M onde $a = b$ e U onde $a \neq b$. O conjunto M é o conjunto de verdadeiros pares (*matches*), ao passo que U é o conjunto de falsos pares (*unmatched*). O problema é determinar a qual conjunto um par de registros pertencerá. Na realidade, um terceiro conjunto P , de possíveis pares (*possibly matched*), é introduzido para acomodar essas situações onde o estado de uma ligação de um par de registros não pode ser definido apenas com base na informação disponível a partir das bases de dados envolvidas. Se um par de registros é atribuído a P , um especialista – um juiz – deve manualmente examinar o par. Assumiremos que o especialista pode sempre identificar o estado correto de uma ligação (M ou U) de cada par de registros com ou sem informação extra.

Assuma que n campos comuns, $f_1, f_2, ..., f_n$ de cada registro da base de dados A e B são escolhidos para comparação. Para cada par de registros $r_{i,j} = (r_i, r_j)$, a comparação de todos campos resulta em um vetor de n valores, $c_{i,j} = \left[c_1^{i,j}, c_2^{i,j}, ..., c_n^{i,j} \right]$ onde $c_k^{i,j} = C_k \left(r_i, f_k, r_j, f_k \right)$ e C_k é a função de comparação que comprara os valores do campo do registro f_k . O vetor, $c_{i,j}$, é chamado vetor de comparação e o conjunto de todas os vetores de comparação é chamado espaço de comparação. Uma função de comparação C_k é um mapa do produto cartesiano do(s) domínio(s), D_k , para o campo f_k para um domínio de comparação R_k ; formalmente, $C_k : D_k \to R_k$. Um exemplo simples de função de comparação é

$$C_I \left(v_1, v_2 \right) = \begin{cases} 0 \text{ se } v_1 = v_2 \\ 1 \text{ caso contrário} \end{cases} \tag{0.6}$$

onde $R_I = \{0,1\}$. O valor computado por C_I é chamado valor de comparação binária. Outros dois tipos adicionais de valores de comparação produzidos por funções de comparação são os categóricos e os contínuos.

A finalidade de um modelo de decisão é determinar o estado da ligação de um par de registros dado seu vetor de comparação $c_{i,j}$. Os modelos de decisão variam em complexidade, diversas propostas nesse sentido foram apresentadas nos últimos tempos. As principais variações vão do tipo de valores de comparação utilizados até mesmo a casos em que alguma quantidade de dado é necessário para treinamento do modelo(Fellegi and Sunter, 1969).

8.7.1 Modelo probabilístico baseado em erro

O modelo probabilístico baseado em erro foi definido por Fellegi e Sunter(Fellegi and Sunter, 1969) e atribui um peso $w_k^{j,j}$ para cada par de registros formado pelos i-ésimo e j-ésimo registros. Formalmente temos a seguinte definição:

$$w_k^{j,j} = \begin{cases} \log \left(m_k / u_k \right) & \text{se } c_k^{i,j} = 0 \\ \log \left(\left(1 - m_k \right) / \left(1 - u_k \right) \right) & \text{se } c_k^{i,j} = 1 \end{cases} \tag{0.7}$$

Onde m_k e u_k são probabilidades condicionais de observar que dois valores do campo k são iguais dado que o par de registro $r_{i,j}$ é uma ligação verdadeira (*true match*) e uma verdadeira não ligação respectivamente. Matematicamente, isso pode ser definido como segue:

$$\begin{aligned} m_k &= \text{Prob}\left(c_k^{i,j} = 0 \mid r_{i,j} \in M\right) \\ u_k &= \text{Prob}\left(c_k^{i,j} = 0 \mid r_{i,j} \in U\right) \end{aligned} \tag{0.8}$$

Devemos considerar que a ponderação $w_k^{j,j}$ é grande (positiva) para um par ligado e pequena (negativa) para um par não ligado. Uma decisão é feita para cada par de registro ao se calcular a ponderação composta $W\left(r_{i,j}\right) = \sum_{k=1}^{n} w_k^{i,j}$ e ao se comparar esse valor contra dois valores limítrofes t_1 e t_2, onde $t_1 < t_2$. De forma específica, a decisão é feita pelos seguintes critério:

$$\begin{aligned} &r_{i,j} \in M \text{ se } W\left(r_{i,j}\right) \geq t_2 \\ &r_{i,j} \in U \text{ se } W\left(r_{i,j}\right) \leq t_1 \\ &r_{i,j} \in P \text{ se } t_1 \leq W\left(r_{i,j}\right) \leq t_2 \end{aligned} \tag{0.9}$$

Em simplificação para ao leitor, se $W\left(r_{i,j}\right) \geq t_2$ o par de registros é considerado uma ligação; se $W\left(r_{i,j}\right) \leq t_1$ o par de registros é classificado como não ligação e se $W\left(r_{i,j}\right)$ estiver entre t_1 e t_2 ele é uma provável ligação. Lembremo-nos de que realizaremos ainda uma decisão de aceitar ou não o par dentro dessa probabilidade de ser uma ligação. O principal propósito deste modelo é portanto determinar estimativas das probabilidades condicionais m_k e u_k para $k = 1, 2, ..., n$, e também estimativas para os dois valores limítrofes t_1 e t_2.

8.7.2 Modelo probabilístico baseado em custo

No modelo anteriormente apresentado, os valores limítrofes para a tomada de decisão eram obtidos pela minimização da probabilidade de tomar uma

decisão errada em relação ao estado da ligação de um par de registros. Ao se adotar essa metodologia, estamos aceitando implicitamente que todo e qualquer tipo de erro é igualmente custoso – danoso. Em verdade, na maioria dos problemas, isso acontece raramente. A minimização da probabilidade de erro não é o melhor critério a ser usado na definição de uma regra de decisão, pois a classificação errada de diferentes pares de registros podem ter diferentes consequências. Por exemplo, quando pareamos uma base de dados do registro de câncer com a base de dados de um hospital, uma ligação perdida resulta em um paciente com câncer perdido no seguimento de um estudo, ou testes de seguimento. Isso tem uma consequência muito séria, ao passo que uma falsa ligação formada resulta em um paciente sem câncer passar a ser avaliado.

Uma proposta de modelo de decisão que minimiza o custo da tomada de decisão é proposto por Verykios et al. (Verykios, Moustakides, and Elfeky, 2003). Esse modelo usa um modelo bayesiano de custo constante de erro para derivar a regra de decisão para uma dada matriz de custo. Para o pareamento de registros, a matriz de custo D tem dimensão 3×2. Consideremos d_{ij} o custo de adotar a decisão i quando o par de registros a ser comparado corresponder a uma das três regiões, dada por uma regra de decisão no espaço de decisão – onde M região das ligações, P as possíveis ligações e U das não-ligações. Assim j se refere ao real estado da ligação, isto é, M' e J'. A regra de decisão é então obtida pela minimização do custo médio $\overline{d}$, que pode ser expresso da seguinte maneira:

$$\begin{aligned}\overline{d} = {} & d_{MM'}\cdot\mathrm{Prob}\left(M,M'\right)+d_{MU'}\cdot\mathrm{Prob}\left(M,U'\right)+\\ & d_{PM'}\cdot\mathrm{Prob}\left(P,M'\right)+d_{PU'}\cdot\mathrm{Prob}\left(P,U'\right)+\\ & d_{UM'}\cdot\mathrm{Prob}\left(U,M'\right)+d_{UU'}\cdot\mathrm{Prob}\left(U,U'\right)\end{aligned} \qquad (0.10)$$

onde $\mathrm{Prob}(i,j)$ denota a probabilidade conjunta de que a decisão i seja adotada quando o real estado da ligação seja j. Com base no teorema de Bayes e substituindo as probabilidades acima pelas probabilidades a priori de M' e U', e as densidades de probabilidade do vetores de comparação dados os estados das ligações, a equação acima pode ser sumarizada por uma regra de decisão similar ao modelo de decisão baseado em erro. Em resumo, a diferença entre ambos modelos apresentados é que os valores limítrofes das regiões dependem da matriz de custo adotada.

8.7.3 Modelo de decisão baseado em aprendizagem indutiva

Uma das limitações dos modelos probabilísticos acima é que eles não podem lidar muito bem com de vetores de comparações a medidas contínuas ou numéricas. Os modelos baseados em técnicas de aprendizagem automática, ou aprendizado de máquina, podem superar essas dificuldades. Um desses modelos de decisão é o baseado em aprendizagem indutiva, capaz de lidar com todos os tipos de vetores de comparação(Elfeky, Verykios, and Elmagarmid, 2002).

Em aprendizado indutivo, um conjunto de padrões de treinamento, em que a classe de cada padrão é conhecida a priori, é usado para construir um modelo que possa ser usado posteriormente para predizer a classe de cada padrão não classificado. Uma instância de treinamento tem a forma $< x, f(x) >$, onde x é um padrão e $f(x)$ é uma função de valores discretos que representa a classe de padrões x - ou seja, $f(x) \in L_1, L_2, ..., L_l$ onde l é o número de possíveis classes. No caso do pareamento de registros, $_x$ é o vetor de comparação c , l é $2(M \text{ e } U)$ e $f(c)$ é o estado da ligação correspondente – de forma que $f(c) \in M, U$. Uma das técnicas mais populares de classificação é a árvore de decisão, usada para explorar as regularidades existentes entre as observações usadas como dados de treinamento. As predições são realizadas com base na similaridade de situações previamente encontradas. A precisão desse tipo de modelo de decisão depende fundamentalmente da representatividade dos dados de treinamento – ou seja, se os dados de treinamento não representam adequadamente os dados posteriormente avaliados, os resultados não serão de boa qualidade.

8.7.4 Modelo de decisão baseado em conglomerados

Uma das limitações existentes nos modelos de decisão baseados em aprendizado indutivo é que eles supõem a existência de dados para treinamento do modelo. Isso nem sempre é uma realidade. Além disso, métodos de aprendizado não

supervisionados, como o modelo baseado em conglomerados, têm ganhado força nos últimos tempos entre pesquisadores de pareamento de dados. Sua popularidade crescente se justifica por não necessitar de dados para treinamento do modelo. Elfeky et al. (Elfeky, Verykios, and Elmagarmid, 2002) utilizam agrupamento em conglomerado de k-médias (*k-means*) para agrupar os pares de registros em três conglomerados – ligações, não-ligações e possíveis ligações.

Uma vez computados os clusters, os pares de registros se encontram classificados sem a necessidade de qualquer outra espécie de intervenção. Porém esta é apenas uma primeira etapa, pois os pares de registros alocados no conglomerado das possíveis ligações. A segunda etapa de separação depende da suposta formação de uma distribuição bimodal dos valores de comparação. Nessa etapa então será necessário determinar um ponto de corte para assumir que um conjunto de pares de registros é ligação ou não. Naturalmente que isso é bastante não desejado na maioria dos problemas do mundo real, pois uma revisão manual (*clerical review*) é bastante custosa e talvez ineficiente em regiões cinza – regiões onde distinguir o verdadeiro do falso é bastante complicado.

8.8 Definição da classificação de pares

8.8.1 Inspeção manual de pares (clerical review)

Os modelos de decisão aplicados ao pareamento de registros probabilístico terminam por formar regiões de decisão sobre o espaço dos escores de similaridade dos pares, como mostra a FIGURA 6: região de aceitação (além do ponto crítico superior), região de rejeição (aquém do ponto crítico inferior) e uma região de prováveis pares. Esta última região – também conhecida como região cinzenta, pois nela há dúvidas sobre a real condição do par de registros – pode ser destinada a uma revisão manual por um humano.

Naturalmente que ao adotar esse método de classificação dos pares, estamos assumindo que um humano tem capacidade de discernir o verdadeiro do falso. Inerente ao humano está o erro, assim essas classificações podem levar a erros de classificação. Geralmente em pareamento de dados altos escores de similaridade (100% conforme exemplo da FIGURA 6) estão associados a verdadeiros pares e baixos escores a não pares. Ao observar a distribuição dos escores de todos os pares

do pareamento de dados, observaremos a mistura da curva dos escores dos verdadeiros pares (a direita) e a curva dos escores dos falsos pares (a esquerda). Com grande frequência, se bem ajustado o modelo de pareamento de dados, observarmos máximos (moda) locais em escore 100% de similaridade (ou valor nominal equivalente) e outro máximo local em escore 0%. A FIGURA 6 é apenas uma ilustração simplificada do comportamento dessas curvas, posto que em casos práticos, seus aspectos revelam bastante irregularidades – não suavidade.

Esta região, em geral, apresenta maior probabilidade de se cometer os erros de classificação: erro de falso positivo (o par de registros não pertence à mesma entidade e o classificamos como par da mesma entidade, uma ligação ou match), ou erro de falso negativo (o par de registros pertence à mesma entidade e o classificamos como par de entidades distintas, não ligação, não match).

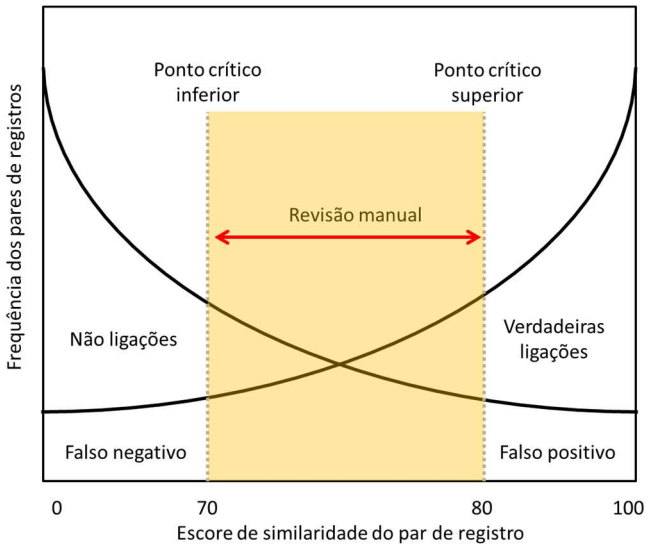


FIGURA 6 – Representação típica dos escores do pareamento para revisão manual.

Nota-se que a revisão manual pode gerar pares que são falsos positivos e falsos negativos, além daqueles que decorrem do algoritmo implementado pelo modelo de decisão. Entretanto não há, entre os problemas reais, como aferir o erro cometido pela revisão manual (Grannis, Overhage, Hui, and McDonald, 2003).

8.8.2 Detecção do ponto de corte por inspeção automática

Grannis e colaboradores (Grannis, Overhage, Hui, and McDonald, 2003) desenvolveram uma metodologia que utiliza o pareamento de registros

probabilístico sem revisão manual humana. Eles empregaram uma função de estimador baseada no algoritmo de Maximização da Esperança (*Expectation Maximization* – EM) que tem a finalidade de determinar um único ponto de corte para pares considerados ligações (Grannis, Overhage, Hui, and McDonald, 2003) – valor limítrofe para separação da região U (não ligações) e região M (ligações), dentro da região P (prováveis ligações) – vide ponto de corte único em FIGURA 7.

Ao comparar essa metodologia não supervisionada com os resultados do método que tem a revisão manual como padrão ouro, chegaram a resultados como 99.95% de especificidade e 99,95% de sensibilidade. Evidenciando assim a superioridade do método EM sobre algoritmos de vinculação de bases determinísticos.

A metodologia sem supervisão humana se baseia no modelo de decisão de Felligi-Sunter, método de pareamento de registros probabilístico. A estrutura das regiões de classificação dos pares de registros se altera com a aplicação dessa metodologia, representada na figura abaixo:

O objetivo do problema é classificar cada par de registro apenas como ligação ou não-ligação. Matematicamente isso é dado pela soma das componentes de peso de cada um dos identificadores usados no pareamento – para o j-ésimo par de registro temos que

$$\text{Score} = \sum_{k=1}^{n} \log\left(\frac{m_k}{u_k}\right)^{\gamma_k^i} \log\left(\frac{1-m_k}{1-u_k}\right)^{1-\gamma_k^i} \tag{0.11}$$

onde para o k-ésimo identificador no j-ésimo par de registro:

n é o número do identificador por registro;

γ_k^i o valor de concordância/discordância observada (1=concorda;0=discorda);

m_k é a estimativa da taxa de concordância do identificador entre as verdadeiras ligações;

u_k é a estimativa da taxa de concordância do identificador entre as falsas ligações.

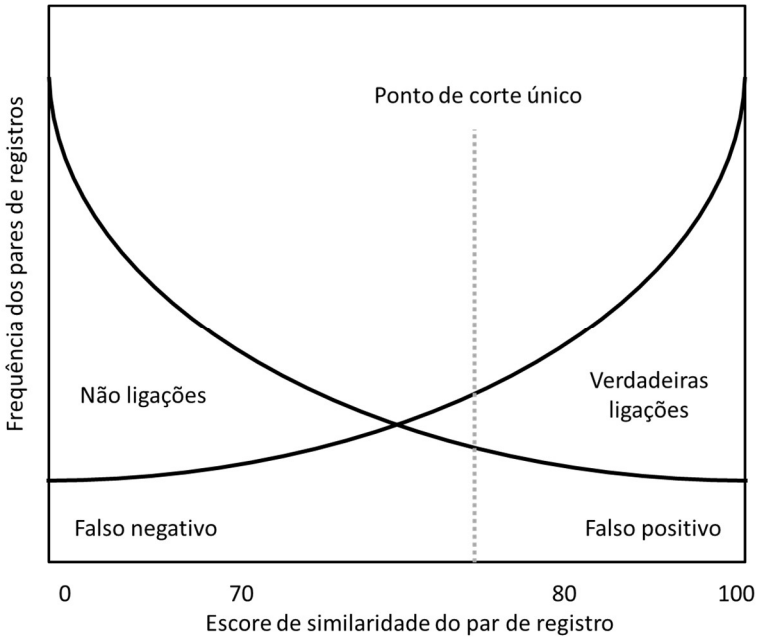


FIGURA 7 – Ponto de corte do escore único para classificação de pares de registros.

Como não são conhecidos os verdadeiros estados dos pares – ou seja, quais pares são verdadeiros ou falsos – os valores de m_k e u_k são estimados a partir dos dados. A etapa de maximização é feita ao se ajustar o logaritmo da função de máxima verossimilhança para o pareamento probabilístico de registros para zero. Então realizar as estimativas dos parâmetros desconhecidos. O logaritmo da função de máxima verossimilhança para o pareamento probabilístico de registros é

$$\ln f\left(d\,|\theta\right)=\sum_{j=1}^{N} g_j \left[\ln P\left(\gamma^j\,|M\right),\ln P\left(\gamma^j\,|U\right)\right]^T+\sum_{j=1}^{N} g_j \left[\ln p,\ln\left(1-p\right)\right]^T \quad (0.12)$$

onde d é observado incompleto dado;

θ é o conjunto de parâmetros (m_1 ,m_2 ,..., m_n , $u_1,u_2,...,u_n,p$);

N número total de pares de registros;

g_j (1,0) para pares ligados e (0,1) para pares não ligados;

γ^j vetor de concordância/discordância do identificador observado para o j-ésimo par de registro;

p a proporção de pares que verdadeiramente são ligações.

A etapa da esperança é dada pelos valores desconhecidos g_j , que são estimados usando $\left(g_m\left(\gamma^j\right), g_u\left(\gamma^j\right)\right)$ onde

$$g_m\left(\gamma^j\right) = \frac{p\prod_{k=1}^{n} m_k \gamma_k^j \left(1-m_k\right)^{1-\gamma_k^j}}{p\prod_{k=1}^{n} m_k \gamma_k^j \left(1-m_k\right)^{1-\gamma_k^j} + \left(1-p\right) p\prod_{k=1}^{n} u_k \gamma_k^j \left(1-u_k\right)^{1-\gamma_k^j}} \quad (0.13)$$

onde $g_u\left(\gamma^j\right)$ é a similaridade derivada.

Para a etapa de maximização, as derivadas parciais para cada um dos três problemas de maximização são ajustados para zero, em relação a equação para os parâmetros desconhecidos. Dessa forma temos que

$$p = \frac{\sum_{j=1}^{N} g_m\left(\gamma^j\right)}{N} \quad (0.14)$$

$$m_k = \frac{\sum_{j=1}^{N} \gamma_k^j \square g_m\left(\gamma^j\right)}{\sum_{j=1}^{N} g_m\left(\gamma^j\right)} \quad (0.15)$$

$$u_k = \frac{\sum_{j=1}^{N} \gamma_k^j \square g_u\left(\gamma^j\right)}{\sum_{j=1}^{N} g_u\left(\gamma^j\right)} \quad (0.16)$$

O processo de estimação então computa os escores para conjuntos de 6.000 pares de registros usando a equação (0.11). É assumida a independência condicional entre os identificadores, então são computadas as estimativas para taxas de verdadeiros positivos e de verdadeiros negativos para cada vetor de comparação $\left(\gamma^j\right)$ usando

$$TP\left(\gamma^j\right) = P\left(\gamma^j \middle| M\right) = \prod_{k=1}^{n} m_k^{\gamma_k^j} \left(1-m_k\right)^{1-\gamma_k^j} \quad (0.17)$$

$$TN\left(\gamma^j\right) = P\left(\gamma^j \middle| U\right) = \prod_{k=1}^{n} u_k^{\gamma_k^j} \left(1-u_k\right)^{1-\gamma_k^j} \quad (0.18)$$

onde $TP(.)$ são verdadeiros positivos e $TN(.)$ são verdadeiros negativos. Assim para um conjunto de dados com n identificadores por registro, existem $2^{(n)}$ únicos vetores de comparação de concordância/discordância $\left(\gamma^{j}\right)$. As estimativas para taxas de verdadeiros positivos e negativos são computadas para cada um dos vetores únicos segundo as equações (0.17) e (0.18). Os vetores são então ordenados ascendentemente pelo escore e são computadas as estimativas de sensibilidade e especificidade como uma função dos escores dos pares de registros. As estimativas de sensibilidade e especificidade são comparadas sob vários pontos de corte para escore dados pela função de máxima verossimilhança, sendo selecionado a solução ótima(Grannis, Overhage, Hui, and McDonald, 2003).

Em resumo, usando as informações dos dados observados, através da estimação da proporção de verdadeiros e falsos pares de registros, usando a função de máxima verossimilhança para o modelo probabilístico de Felligi-Sunter, o método busca encontrar um ponto de corte que minimiza o número de falsos positivos e falsos negativos – que termina por ser também uma maximização da sensibilidade e especificidade (Leicester Gill, 2001; Grannis, Overhage, Hui, and McDonald, 2003).

É importante destacar que o algoritmo de Maximização da Esperança, do ponto de vista teórico, considera que exista uma completitude perfeita das variáveis do pareamento de dados. O algoritmo tem a sua capacidade de estimação das probabilidades *u* e *m* bastante comprometidas a medida em que o volume de valores faltantes (*missing values*) aumenta nas variáveis do pareamento de dados (Bauman, 2006; Durham, 2012)

8.9 Aplicações do pareamento de registros no SUS

Hoje podemos observar que o pareamento de registros tem sido crescentemente incorporado ao cotidiano de pesquisadores, gestores de saúde, e gerentes de sistemas de informação. A técnica é utilizada para responder perguntas de interesse da saúde pública (vigilância em saúde). Classificadas como aprimoramento da qualidade e eliminação de duplicidades temos como exemplos: esforços de redução de causas de óbito mal definidas (Teixeira, Bloch, Klein, and Coeli, 2006), qualificação da informação sobre mortalidade infantil (Mendes, Lima, Sá, Oliveira,

and Maia, 2012), identificar subnotificação de casos de tuberculose no Brasil (Oliveira, Pinheiro, Coeli, Barreira, and Codenotti, 2012).

O pareamento de registros pode ser aplicado para estimativa de registros sobre um determinado evento (ou populações) – por exemplo, impacto que casos identificados de subnotificação para tuberculose causariam na taxa de notificação de casos novos e proporção de óbitos (Oliveira, Pinheiro, Coeli, Barreira, and Codenotti, 2012), por meio de técnicas de captura e recaptura (extremamente comum no caso da AIDS, estimativa do número de casos de leptospirose sintomática na região central do Rio Grande do Sul (Brum, 2005), estimativas de sub-registro de leptospirose humana (Brum and Kupek, 2005)). Também é aplicado a reidentificação/anomização de registros visando a proteção da privacidade dos cidadãos mencionados nas fontes de dados (e/ou a de seus familiares). Encontra aplicação em Epidemiologia (orientada para pacientes e estudos longitudinais) (Achimugu Philip, n.d.), estatísticas do Censo (Jaro, 1989; William E. Winkler, 1997), listas de contatos comerciais (limpeza e atualização), detecção de crimes e fraudes. Em estudos epidemiológicos, o pareamento de registros é frequentemente utilizado como o passo inicial para a análise de dados, organização das informações em série histórica(Moura, Prestes, Duncan, and Schmidt, 2014) e/ou projetos de mineração de dados.

Várias perguntas da saúde pública demandam consultas a bases de dados oriundas de vários sistemas, integradas especialmente para responder a estas perguntas. No Brasil, observamos hoje um grande esforço do Ministério da Saúde, via Secretaria de Vigilância de Saúde (SVS). A SVS vem utilizando cada vez mais das técnicas de vinculação de bancos de dados, gerando produtos de importância para aprimorar a qualidade e completitude das informações e, consequentemente a gestão do Sistema de Saúde. São exemplos de trabalhos da SVS usando pareamento de registros.

Projeto de aprimoramento da qualidade da informação sobre a causa básica dos óbitos e aumento da cobertura do SIM e SINASC. A partir do relacionamento SIM x SIH, SIM x APAC, este projeto vem contribuindo de forma inequívoca com a investigação de óbitos e o resgate da causa de óbitos para registros que se encontravam na base de mortalidade com causas mal definidas. Para o Brasil,

o percentual de causas definidas aumentou de menos de 85% em 2000 para 94% em 2013, apresentando resultados bastante expressivos nas regiões norte (passando de 76% para 89%) e nordeste (passando de 72% para 93%), com resultados ainda mais impressionantes na análise do recorte por UF (Filho et al., 2016).

A partir do relacionamento SIM x Registro Civil, e SINASC x Registro Civil, o projeto tem orientado também a busca ativa de eventos que se encontram fora das bases de mortalidade e nascidos vivos, e deverá contribuir para o aumento da cobertura dos dois sistemas.

Deduplicação da base de Tuberculose do SINAN, que mostra em seus resultados que de 2000 a 2004, 73,7% das notificações eram únicas, 18,9% formavam duplas; 4,7% eram triplas e 2,7% grupos de quatro ou mais registros. Do total de registros repetidos, 47,3% - transferência entre unidades de saúde, 23,6% reingresso, 16,4% duplicidade verdadeira, 10% recidiva, 2,5% foram inconclusivos e 0,2% tinham dados faltando em variáveis específicas. A exclusão de registros indevidamente repetidos resultou em redução na taxa de incidência por 100 mil habitantes de 6,1% em 2000 (de 44 para 41,3), 8,3% em 2001 (de 44,5 para 40,8), 9,4% em 2002 (de 45,8 para 41,5), 9,2% em 2003 (de 46,9 para 42,6) e 8,4% em 2004 (de 45,4 para 41,6). Este tipo de ação aumenta a precisão dos diagnósticos de situação de saúde e a qualificação impacta substancialmente na definição de linhas de base para monitoramento do impacto de das políticas que visam controlar.

O pareamento de registros pode ser utilizado também para verificar a qualidade de registros de base populacional e/ou sistemas. Um exemplo, no Brasil, é o trabalho de Oliveira e colaboradores 2014 para o câncer, que avaliou a confiabilidade das causas básicas de óbito por câncer por meio do relacionamento probabilístico entre Sistema de Informações sobre Mortalidade (SIM) e Registro de Câncer de Base Populacional (RCBP) de Goiânia, Goiás, Brasil, entre 2000 e 2005. O pareamento possibilitou a reclassificação de 67% de câncer de localização mal definida no SIM usando as informações do RCBP. O estudo foi muito positivo no sentido de qualificar as estimativas de mortalidade por câncer em áreas cobertas por RCBP(Oliveira, Silva, Curado, Malta, and Moura, 2014).

O pareamento de dados ainda pode ser utilizado para o acompanhamento e obtenção de informações para avaliação do impacto de

intervenções governamentais para o desenvolvimento social. Com uma coorte de 103 milhões de pessoas, mais precisamente, todos indivíduos que receberam pelo menos um pagamento do Programa Bolsa Família entre 2007 e 2012, a técnica possibilitou aos pesquisadores avaliar o impacto do Programa Bolsa Família (Campello and Neri, 2013) sobre, entre outros, incidência de doenças como lepra, AIDS-HIV e tuberculose (Rasella, Aquino, Santos, Paes-Sousa, and Barreto, 2013; Torrens, 2015; Torrens et al., 2016). O grupo formado por uma parceria entre as universidades federais da Bahia e Brasília, London School of Hygiene & Tropical Medicine; apoio do Ministério da Saúde, Ministério do Desenvolvimento Social, realizou o pareamento dos dados do Cadastro Único (CADU), uma base de dados socioeconomica usada para registro de pessoas beneficiárias de programas socias do Governo Federal, com as bases do SUS de SIH, SINAN e SIM. Os resultados foram bastante satisfatórios, atingindo mais de 95% de verdadeiros positivos entre todos os pareamentos e, para o pareamento entre CADU e SIM, 100% de sensibilidade, especificidade e valor preditivo positivo (Pita et al., 2015; Barreto et al., 2017) .

8.10 Ferramentas computacionais para pareamento de registros

8.10.1 FRIL

O FRIL é uma ferramenta computacional, de código fonte livre e aberto escrito em Java, capaz de realizar o pareamento de registros de forma facilitada. Ele é o resultado de um projeto conjunto entre a Emory *University* e o *Centers for Disease Control and Prevention* (CDC).

O FRIL se destaca de outros programas de computador de sua categoria por oferecer ao usuário a possibilidade de configurar rapidamente, e de forma simplificada, os parâmetros envolvidos no pareamento das bases de dados. O fluxo de trabalho para o pareamento é ordenado e sequencial (FIGURA 8) – o que funciona como um facilitador para usuários não tão avançados. Mesmo configurações avançadas do programa como número de processadores e núcleos a serem utilizados; quantidade de memória RAM alocada; e paginamento de memória (*memory paging*) em disco, são realizadas visualmente.

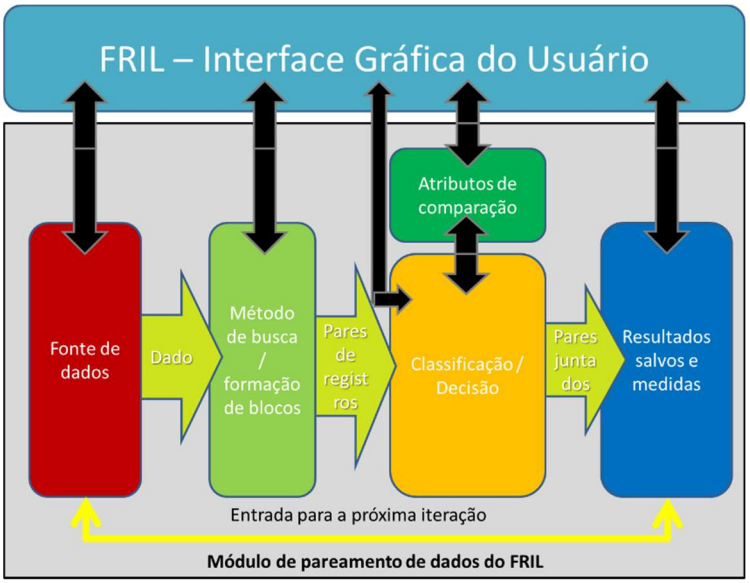


FIGURA 8 – Fluxo de Trabalho no FRIL e relacionamento de suas camadas de interfaceamento.

Em sua primeira etapa, o FRIL apresenta janelas onde o usuário define as duas bases de dados que serão pareadas. Um intuitivo e rico conjunto de ações avançadas são oferecidas para a transformação e reconciliação das informações nas bases e entre bases – modificação de *schema*/data. Por exemplo, uma variável da base A pode apresentar data de nascimento no formato DDMMAAAA, enquanto na base B essa informação se encontra no formato MMDDAAAA. Problema esse que pode ser solucionado aplicando uma função de reorganização dos caracteres dada pelo FRIL.

A quantidade de métodos de busca oferecidos pelo FRIL cobre hoje os métodos convencionais e outros tidos como experimentais. Entre esses métodos estão *sorted neighborhood method*, *blocking method*, *nested loop join*.

A fase para a definição de critérios de classificação e decisão do pareamento de registros permite que o usuário possa configurar as variáveis a serem utilizadas na vinculação das bases e o ajuste fino dos parâmetros para as métricas de similaridade. Todo processo do ajuste de parâmetros pode ser inspecionado pelo usuário através de uma análise dinâmica da performance dos parâmetros. Isto é, definido o valor para um parâmetro, o usuário pode solicitar que uma nova janela seja aberta e o pareamento da variável em configuração seja exibido em tempo real.

Outro diferencial do FRIL é o suporte transparente ao usuário para o processamento da tarefa de pareamento em máquinas com processadores de múltiplos

núcleos – o usuário pode definir o número de núcleos que deseja alocar para o pareamento, ficando a cargo do FRIL determinar a melhor forma de distribuir o processamento entres os núcleos. Toda essa distribuição entre núcleos acontece sem demandar trabalho/custo algum ao usuário.

O fato do FRIL ter sido escrito totalmente em Java, o torna apto a ser executado nos principais sistemas operacionais, por exemplo, Unix, Linux, Windows e MacOS (Jurczyk, Lu, Xiong, Cragan, and Correa, 2008; FRIL: Fine-grained record linkage software [Internet], 2011).

Uma outra grande implementação oferecida pelo FRIL é a ferramenta open-source, LinkIT. Essa ferramenta tem a finalidade de preservar a privacidade de registro durante o pareamento (privacy preserving record Linkage) e integração dos sistemas via transformações de dados. O LinkIT implementa novos algoritmos que suportam transformações dos dados envolvendo informações sensíveis (Bonomi, Xiong, and Lu, 2013).

8.10.2 Parea

O PAREA, desenvolvido por Walter dos Santos Filho na Universidade Federal de Minas Gerais, é uma ferramenta que realiza o pareamento probabilístico de registros focada na paralelização do processo, o qual, segundo seus autores se torna ainda mais necessária quando é necessário trabalhar com bases de dados reais. Para diminuir o tempo de processamento, a ferramenta utiliza formas de paralelizar o algoritmo de pareamento de registro, oferecendo uma solução capaz de escalar bem quando executada em um cluster de computadores. Seus autores afirmam que a ferramenta não demanda uma grande quantidade de recursos, especialmente memória principal – realizando o uso de cache de comunicação, recurso fundamental para a escalabilidade. O PAREA, evolução da ferramenta FERAPARDA, pode ser usada em diferentes bases de dados, desde bases comerciais até bases da saúde e de programas sociais a fim de melhorar a qualidade da informação e melhorar a qualidade dos serviços que se baseiam em tal informação (Filho, 2008).

8.10.3 Reclink

O Reclink é um sistema de relacionamento de bases de dados fundamentado na técnica de relacionamento probabilístico de registros, desenvolvido na linguagem C++ com o ambiente de programação Borland C++ Builder versão 3.0. Ele é o produto de um projeto liderado pelos professores Kenneth R. de Camargo Jr. e Cláudia M. Coeli.

Uma variação recente do Reclink, conhecido como Open Reclink, é um esforço que segue a filosofia do software livre, oferecendo o seu código fonte aberto e de livre distribuição.

O Reclink oferece ao usuário facilidades para a padronização das informações da base de dados usadas no pareamento de registros. Transformações para primeiro nome são oferecias conforme pesquisas para nomes brasileiros:

Se a primeira letra é W e a segunda A, substitui a primeira por V;

Se a primeira letra é H, exclui a primeira letra;

Se a primeira letra é K e a segunda A, O ou U, substitui a primeira por C;

Se a primeira letra é Y, então é substituída por I;

Se a primeira letra é C e a segunda é E ou I, então a primeira letra é trocada por S;

E se a primeira letra é G e a segunda é E ou I, então a primeira letra é substituída por J.

O Reclink automatiza todos as etapas do processo de pareamento de registros. Uma intervenção manual é necessária apenas para a revisão manual dos pares. O programa utiliza o padrão xBase (extensão DBF) e o conjunto de configurações definidas pelo usuário podem ser salvos em arquivos com a extensão RSP. A organização da blocagem pode ser exportada para arquivos com extensão RSD. Arquivos contendo os parâmetros para a geração de arquivos combinados usam a extensão RSC.

A etapa do pareamento utiliza escores finais probabilísticos, obtidos através dos fatores de ponderação de concordância e discordância totalmente definidos pelo usuário – a partir desses escores acontece a classificação dos pares relacionados.

O programa permite ao usuário o uso de uma ou mais variáveis das bases de dados para a realização da vinculação. O resultado é dado em um novo arquivo contendo as informações previamente determinadas pelo usuário para compor a saída. Dessa forma, algumas variáveis podem ser usadas no pareamento, mas a saída do programa ser bem mais enxuta – sem essas variáveis.

Uma documentação sobre o programa está disponível em seu manual bastante detalhado. Com isso o Reclink é uma boa ferramenta para pesquisadores ou profissionais que estão iniciando suas empresas em pareamento de dados (Jr, De, and Coeli, 2000).

8.10.4 Link King

O Link King é uma ferramenta para o pareamento de registros de domínio público, implementado em SAS. Ele se encontra disponível gratuitamente em http://www.the-link-king.com. O algoritmo do Link King foi pensado para ser flexível, aplicável aos mais comuns problemas de vinculação de bases de dados, mas sempre com grande preocupação quanto ao consumo de recursos computacionais. Essencialmente três métodos de pareamento estão disponíveis neste programa, sendo dois determinísticos e um probabilístico. Todos métodos implementados com valor preditivo positivo superior a 90% segundo pesquisas e sensibilidade na casa dos 80% (ligações-vinculação de bases) e 90% (método probabilístico) (Gomatam, Carter, Ariet, and Mitchell, 2002; Grannis, Overhage, and McDonald, 2002; Campbell, Deck, and Krupski, 2008; Bessel, 2010; Dedupe Software / Record Linkage Software by The Link King FREE ! [Internet], n.d.).

8.10.5 Bibliotecas Python para pareamento de registros

Entre outras bibliotecas escritas em Python, destaca-se pela riqueza em sua implementação a Febrl (Christen, 2008*a*). Seu código é integralmente escrito em linguagem de script orientada a objetos Python e distribuída como software livre de código fonte aberto. Essa biblioteca é organizada com caráter modular e flexível, objetivando fácil utilização, modificação e incorporação dentro de outros programas escritos em Python – caso o usuário decida escrever suas próprias implementações.

Entre as técnicas e métodos implementados na Febrl podemos destacar os métodos de bloqueamento padrão e indexação de busca bigram. Outros

métodos como o *sorted neighbourhood* e *canopy clustering* com TFIDF (*Term Frequency/Inverse Document Frequency*).

A Febrl oferece condições ideais para a implementação de novos algoritmos, técnicas e métodos no que se refere a limpeza de dados (*data cleaning*), padronização de informação e pareamento de registros (Christen, 2008*a*).

8.10.6 Pacotes R para pareamento de registros

Entre as implementações de pareamento de registros no R, pode-se destacar o pacote RecordLinkage – a implementação mais completa para esse ambiente estatístico e de programação. O pacote apresenta uma classe de objetos própria para o pareamento de registros, dando ao usuário um grande suporte para desenvolvimento de funções ou métodos particulares – se este não se der por satisfeito já com todas implementações de métricas de similaridade já presentes no R.

O fluxo de trabalho no pacote RecordLinkage é diferente dos outros programas por não apresentar interface gráfica para a manipulação dos objetos de dados usados no pareamento. Todo o processo do pareamento de registros pode ser organizado. O pacote oferece dois bancos de dados como exemplo – estão disponíveis informações como primeiro nome, sobrenome e data de nascimento. O processo de formação dos pares a serem comparados são organizados por duas funções: compare.dedup – para deduplicação de um conjunto de dados; e a compare.linkage – para parear os registros dos dois conjuntos de dados envolvidos no problema. Exemplo de execução dessas funções pode ser visto em QUADRO 3. Caso o usuário deseje usar mais que 2 bancos, o programa realiza o pareamento de dois conjuntos de dados, então pareia o resultado do pareamento anterior com um outro conjunto de dados e assim sucessivamente até concluir a tarefa.

O pacote RecordLinkage disponibiliza para o usuário um conjunto nativo de funções para comparação de *strings*, medidas de similaridade. Entre elas podemos destacar o Soundex para inglês e alemão, Jaro-Winkler, Winkler e Lehvenstein. Medidas de comparação de strings do tipo fuzzy também foram implementadas já normalizadas, isto é, assumem valores entre 0 e 1.

```
> rpairs <- compare.dedup(RLdata500,
+ blockfld = list(1, 5:7),
+ identity = identity.RLdata500)
> rpairs$pairs[c(1:3, 1203:1204), ]
```

```
id1 id2 fname_c1 fname_c2 lname_c1 lname_c2
1 17 119 1 NA 0 NA
2 61 106 1 NA 0 NA
3 61 175 1 NA 0 NA
1203 37 72 0 NA 0 NA
1204 44 339 0 NA 0 NA
by bm bd is_match
1 0 0 0 0
2 0 0 1 0
3 0 0 1 0
1203 1 1 1 1
1204 1 1 1 0
```

QUADRO 3 – Programação em linha de comando para deduplicacão de registros de um conjunto de dados usando o pacote RecordLinkage em R.

O modelo de decisão aplicado no RecordLinkage é o clássico desenvolvido por Fellegi-Sunter (Fellegi and Sunter, 1969), o qual realiza avaliação estocástica e define pontos limítrofes para a classificação dos pares de registros. Quando os pontos limítrofes são definidos, a classificação é realizada com as funções emClassify ou epiClassify – as quais tem como argumentos o objeto do conjunto de dados pareado e suas definições de pontos limítrofes (Sariyar and Borg, 2010; Borg and Sariyar, 2015; BigData.pdf [Internet], n.d.).

OBJETIVOS

1. Objetivos

Objetivo Geral

Gerar base de dados para oncologia, orientada para análise epidemiológica, a partir do pareamento de registros entre as grandes bases de dados do SUS, e utilizá-la para avaliar aspectos relacionados ao tempo decorrido entre diagnóstico e início de tratamento de pacientes com cânceres considerados mais susceptíveis ao tratamento.

Objetivos Específicos

1) Organizar uma base de dados nacional de oncologia (quimioterapia e radioterapia) utilizando pareamento de registros – adaptação da metodologia desenvolvida por Moura et al. 2013 (Lenildo de Moura, 2012; Moura, Prestes, Duncan, and Schmidt, 2014).
2) Descrever a criação da base nacional de oncologia gerada a partir do pareamento de registros da Autorização para Procedimentos de Alta Complexidade (APAC), do Sistema de Informações Hospitalares do SUS (SIH) e do Sistema de Informação sobre Mortalidade (SIM). (Artigo 1)
3) Definir, para cada um dos cânceres potencialmente curáveis, segundo critérios definidos por Farmer et al. 2010, a incidência e a prevalência de casos tratados por quimio- e radioterapia, bem como o tempo de espera entre o diagnóstico e o início de seu tratamento. (Artigo 2)

REFERÊNCIAS BIBLIOGRÁFICAS

Abegunde DO, Mathers CD, Adam T, Ortegon M, Strong K. The burden and costs of chronic diseases in low-income and middle-income countries. The Lancet. 2007;370(9603):1929–38.

Achimugu Philip SA. Modeling and Implementing Record Linkage in Health Information Systems.

Adams C. Dictionary Attack. In: Tilborg HCA van, Jajodia S, editors. Encycl. Cryptogr. Secur. [Internet]. Springer US; 2011 [cited 2015 Feb 17]. p. 332–332. Available from: http://link.springer.com/referenceworkentry/10.1007/978-1-4419-5906-5_74

Alvey W. Record Linkage Techniques - 1997: Proceedings of an International Workshop and Exposition, March 20-21, 1997, Arlington, Va. National Academies; 1997.

Barbosa IR, de Souza DLB, Bernal MM, Costa Í do CC. Cancer mortality in Brazil. Medicine (Baltimore) [Internet]. 2015 Apr 24 [cited 2017 Jul 6];94(16). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4602680/

Barreto M, Alves A, Sena S, Fiaccone R, Amorim L, Ichihara MY, et al. Assessing the accuracy of probabilistic record linkage of social and health databases in the 100 million Brazilian cohort. Int. J. Popul. Data Sci. [Internet]. 2017;1(1). Available from: https://ijpds.org/article/view/276

Bauman GJ. Computation of weights for probabilistic record linkage using the EM algorithm. 2006;

Baxter R, Christen P, Churches T. A Comparison of Fast Blocking Methods for Record Linkage. KDD 2003 Workshop. 2003. p. 25–27.

Bessel M. Linkagem de Dados Utilizando os Programas Link King e SAS® 9.2. 2010 [cited 2015 Jan 25]; Available from: http://www.lume.ufrgs.br/handle/10183/29100

Bittencourt SA, Camacho LAB, Leal M do C. Hospital Information Systems and their application in public health. Cad. Saúde Pública. 2006 Jan;22(1):19–30.

Blakely T, Salmond C. Probabilistic record linkage and a method to calculate the positive predictive value. Int. J. Epidemiol. 2002 Dec;31(6):1246–52.

Boing AF, d'Orsi E, Reibnitz C. Sistema de informações hospitalares do sistema único de saúde (SIH-SUS). Httpsunasusmoodleufscbrfilephp32Contonline14-04un04obj10html [Internet]. 2010 [cited 2015 Feb 8]; Available from: https://ares.unasus.gov.br/acervo/handle/ARES/277

Bonomi L, Xiong L, Chen R, Fung B. Privacy preserving record linkage via grams projections. ArXiv Prepr. ArXiv12082773. 2012;

Bonomi L, Xiong L, Lu JJ. Linkit: privacy preserving record linkage and integration via transformations. Proc. 2013 ACM SIGMOD Int. Conf. Manag. Data. ACM; 2013. p. 1029–32.

Borg A, Sariyar M. RecordLinkage: Record Linkage in R [Internet]. 2015 [cited 2015 Jan 25]. Available from: http://cran.r-project.org/web/packages/RecordLinkage/index.html

Brasil. Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Análise de Situação de Saúde*. Saúde Brasil 2010: uma análise da situação de saúde e de evidências selecionadas de impacto de ações de vigilância em saúde / Health Brazil 2010: health analysis and selected evidences on impact of health surveillance actions. Ministério da Saúde; 2011.

Brasil P. Saúde amplia radioterapia em 22 estados e DF [Internet]. Portal Bras. 2013 [cited 2015 Feb 21]. Available from: http://www.brasil.gov.br/saude/2013/11/saude-amplia-radioterapia-em-22-estados-e-df

Brasil P. Ministério da Saúde expande acesso ao tratamento oncológico [Internet]. Portal Bras. 2014 [cited 2015 Feb 22]. Available from: http://www.brasil.gov.br/saude/2014/01/ministerio-da-saude-expande-acesso-ao-tratamento-oncologico

Brasil P. Ministério da Saúde aprimora oferta de medicamento para câncer no SUS [Internet]. Portal Bras. [cited 2015 Feb 22]. Available from: http://www.brasil.gov.br/saude/2011/03/ministerio-da-saude-aprimora-oferta-de-medicamento-para-cancer-no-sus

Bray F, Jemal A, Grey N, Ferlay J, Forman D. Global cancer transitions according to the Human Development Index (2008–2030): a population-based study. Lancet Oncol. 2012;13(8):790–801.

Brazil, Brazil Congresso Nacional Câmara dos Deputados. Constituição 1988. Brasília, DF, Brasil: Centro de Documentação e Informação, Coordenação de Publicações; 2003.

Brum L, Kupek E. Record linkage and capture-recapture estimates for underreporting of human leptospirosis in a Brazilian health district. Braz. J. Infect. Dis. 2005 Dec;9(6):515–20.

Brum LM. Estimativa do número de casos de leptospirose sintomática na região central do Rio Grande do Sul, Brasil, a partir do método de captura-recaptura. 2005 [cited 2015 Feb 14]; Available from: https://repositorio.ufsc.br/xmlui/handle/123456789/102276

Campbell KM, Deck D, Krupski A. Record linkage software in the public domain: a comparison of Link Plus, The Link King, and a “basic” deterministic algorithm. Health Informatics J. 2008 Mar;14(1):5–15.

Campello T, Neri MC. Programa Bolsa Família: uma década de inclusão e cidadania. 2013;

Casa Civil Brasil. Lei nº 12.732, de 22 de Novembro de 2012. Dispõe sobre O Primeiro Tratamento e Paciente com Neoplasia Maligna Comprovada e Estabelece Prazo Para Seu Início [Internet]. Diário Oficial da República Federativa do Brasil; 2013. Available from: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2012/lei/l12732.htm

Christen P. Febrl -: An Open Source Data Cleaning, Deduplication and Record Linkage System with a Graphical User Interface. Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. [Internet]. New York, NY, USA: ACM; 2008 [cited 2015 Jan 25]. p. 1065–1068. Available from: http://doi.acm.org/10.1145/1401890.1402020

Christen P. Automatic Record Linkage Using Seeded Nearest Neighbour and Support Vector Machine Classification. Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. [Internet]. New York, NY, USA: ACM; 2008 [cited 2015 Jan 25]. p. 151–159. Available from: http://doi.acm.org/10.1145/1401890.1401913

Christen P, Churches T, Hegland M. Febrl – A Parallel Open Source Data Linkage System. In: Dai H, Srikant R, Zhang C, editors. Adv. Knowl. Discov. Data Min. [Internet]. Springer Berlin Heidelberg; 2004 [cited 2015 Feb 8]. p. 638–47. Available from: http://link.springer.com/chapter/10.1007/978-3-540-24775-3_75

Coeli CM, Camargo Jr. KR de. Evaluation of different blocking strategies in probabilistic record linkage. Rev. Bras. Epidemiol. 2002 Aug;5(2):185–96.

Cohen W, Ravikumar P, Fienberg S. A comparison of string metrics for matching names and records. KDD Workshop Data Clean. Object Consol. [Internet]. 2003 [cited 2015 Jan 25]. p. 73–78. Available from: https://www.cs.cmu.edu/afs/cs/Web/People/wcohen/postscript/kdd-2003-match-ws.pdf

Coleman MP, Quaresma M, Berrino F, Lutz J-M, De Angelis R, Capocaccia R, et al. Cancer survival in five continents: a worldwide population-based study (CONCORD). Lancet Oncol. 2008 Aug;9(8):730–56.

Constituição Brasil. LEI Nº 8.080, DE 19 DE SETEMBRO DE 1990 [Internet]. 1990 [cited 2015 Feb 23]. Available from: http://www.planalto.gov.br/ccivil_03/leis/l8080.htm

DATASUS. TabNet Win32 3.0: Mortalidade - Brasil [Internet]. 2017 [cited 2017 Jul 6]. Available from: http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sim/cnv/obt10uf.def

DATASUS Brasil. SIHD - Sistema de Informação Hospitalar [Internet]. [cited 2015 Feb 8]. Available from: http://www2.datasus.gov.br/SIHD/institucional

Durham EA. A framework for accurate, efficient private record linkage. Vanderbilt University Nashville, TN; 2012.

Elfeky MG, Verykios VS, Elmagarmid AK. TAILOR: a record linkage toolbox. 18th Int. Conf. Data Eng. 2002 Proc. 2002. p. 17–28.

Elisangela Santos de Moura. Direito à saúde na Constituição [Internet]. Jus Navig. 2013 [cited 2015 Feb 23]. Available from: http://jus.com.br/artigos/25309/o-direito-a-saude-na-constituicao-federal-de-1988

Estadão. Em cinco anos, gasto com tratamento contra câncer cresceu 66% - Economia - Estadão [Internet]. 2016 [cited 2017 Jul 6]. Available from: http://economia.estadao.com.br/noticias/geral,em-cinco-anos-gasto-com-tratamento-contra-cancer-cresceu-66,10000069529

Estadão. Número de pacientes que recorre ao SUS para tratar câncer aumenta 34% - Saúde [Internet]. Estadão. 2016 [cited 2017 Jul 6]. Available from: http://saude.estadao.com.br/noticias/geral,numero-de-pacientes-que-recorre-ao-sus-para-tratar-cancer-aumenta-34,10000068521

Farmer P, Frenk J, Knaul FM, Shulman LN, Alleyne G, Armstrong L, et al. Expansion of cancer care and control in countries of low and middle income: a call to action. Lancet. 2010 Oct 2;376(9747):1186–93.

Fellegi IP, Sunter AB. A Theory for Record Linkage. J. Am. Stat. Assoc. 1969;64(328):1183–210.

Filho S, Martins A, Cortez-Escalante JJ, França E, Filho S, Martins A, et al. Review of deaths correction methods and quality dimensions of the underlying cause for accidents and violence in Brazil. Ciênc. Amp Saúde Coletiva. 2016 Dec;21(12):3803–18.

Filho W dos S. Algoritmo paralelo e eficiente para o problema de pareamento de dados [Internet]. 2008 [cited 2015 Jan 25]. Available from: http://www.bibliotecadigital.ufmg.br/dspace/handle/1843/RVMR-7L3Q3V

Fitzmaurice C, Allen C, Barber RM, Barregard L, Bhutta ZA, Brenner H, et al. Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life-years for 32 Cancer Groups, 1990 to 2015: A Systematic Analysis for the Global Burden of Disease Study. JAMA Oncol. 2017 Apr 1;3(4):524–48.

Foster Jr B, Evans DP. Comparative q-gram Analysis of Gene Promoter Regions. N. B. 2003;

Freire SM, Gonçalves R de CB, Bandarra AC, Villela MGT, Meire A, Cabral MDB, et al. Análise da efetividade de comparadores de strings para discriminar pares verdadeiros de pares falsos no relacionamento de registro. An. IX Workshop Informática Médica XXIX Congr. Soc. Bras. Comput. Workshop Informática Médica Bento Gonçalves Soc. Bras. Comput. 2009. p. 2119–28.

Furht B, Escalante A. Handbook of Data Intensive Computing. Springer Science & Business Media; 2011.

GBD 2015 Mortality and Causes of Death Collaborators. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980-2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet Lond. Engl. 2016 Oct 8;388(10053):1459–544.

Girianelli VR, Gamarra CJ, Azevedo e Silva G, Girianelli VR, Gamarra CJ, Azevedo e Silva G. Disparities in cervical and breast cancer mortality in Brazil. Rev. Saúde Pública. 2014 Jun;48(3):459–67.

Gomatam S, Carter R, Ariet M, Mitchell G. An empirical comparison of record linkage procedures. Stat. Med. 2002;21(10):1485–96.

Grannis SJ, Overhage JM, Hui S, McDonald CJ. Analysis of a Probabilistic Record Linkage Technique without Human Review. AMIA. Annu. Symp. Proc. 2003;2003:259–63.

Grannis SJ, Overhage JM, McDonald CJ. Analysis of identifier performance using a deterministic linkage algorithm. Proc. AMIA Symp. American Medical Informatics Association; 2002. p. 305.

Gu L, Baxter R. Decision Models for Record Linkage. In: Williams GJ, Simoff SJ, editors. Data Min. [Internet]. Springer Berlin Heidelberg; 2006 [cited 2015 Feb 8]. p. 146–60. Available from: http://link.springer.com/chapter/10.1007/11677437_12

Guerra MR, Bustamante-Teixeira MT, Corrêa CSL, Abreu DMX de, Curado MP, Mooney M, et al. Magnitude and variation of the burden of cancer mortality in Brazil and Federation Units, 1990 and 2015. Rev. Bras. Epidemiol. Braz. J. Epidemiol. 2017 May;20Suppl 01(Suppl 01):102–15.

Hall PA, Dowling GR. Approximate string matching. ACM Comput. Surv. CSUR. 1980;12(4):381–402.

Henrique Hoffmann Monteiro de Castro. Do direito público subjetivo à saúde: conceituação, previsão legal e aplicação na demanda de medicamentos em face do Estado-membro [Internet]. Direito Público Subjetivo À Saúde. 2005 [cited 2015 Feb 23]. Available from: http://www.egov.ufsc.br/portal/sites/default/files/anexos/14736-14737-1-PB.htm

Herzog TN, Scheuren FJ, Winkler WE. Data Quality and Record Linkage Techniques. Springer Science & Business Media; 2007.

IHME. GBD Compare | IHME Viz Hub [Internet]. 2016 [cited 2017 Aug 1]. Available from: http://vizhub.healthdata.org/gbd-compare

INCA. INCA - Agência de notícias -Ministério reafirma compromisso: início de tratamento de câncer em até 60 dias no SUS [Internet]. Minist. Reafirma Compromisso Início Trat. Câncer Em Até 60 Dias No SUS. 2013 [cited 2015 Feb 23]. Available from: http://www2.inca.gov.br/wps/wcm/connect/agencianoticias/site/home/noticias/2013/tempo_de_espera_nao_pode_ultrapassar_sessenta_dias

INCTR The International Network For Cancer Treatment and Research. Cancer in Developing Countries - INCTR – International Network for Cancer Treatment and Research [Internet]. Cancer Dev. Ctries. 2017 [cited 2017 Jul 26]. Available from: http://www.inctr.org/about-inctr/cancer-in-developing-countries/

Instituto Brasileiro de Geografia e Estatística IBGE Brasil. IBGE | Comitê de Estatísticas Sociais | base de dados | metadados | ministério da saúde | sistema de informações ambulatoriais do sus – SIA/SUS [Internet]. [cited 2015 Feb 8]. Available from: http://ces.ibge.gov.br/base-de-

dados/metadados/ministerio-da-saude/sistema-de-informacoes-ambulatoriais-do-sus-sia-sus

Instituto Brasileiro de Geografia e Estatística BRASIL. IBGE | Comitê de Estatísticas Sociais | base de dados | metadados | ministério da saúde | sistema de informações de mortalidade – SIM [Internet]. [cited 2015 Feb 8]. Available from: http://ces.ibge.gov.br/base-de-dados/metadados/ministerio-da-saude/sistema-de-informacoes-de-mortalidade-sim

Instituto Nacional de Alimentação e Nutrição. Portal do Departamento de Atenção Básica - Pesquisa Nacional sobre Saúde e Nutrição (PNSN) [Internet]. 1989 [cited 2015 Feb 22]. Available from: http://dab.saude.gov.br/portaldab/ape_vigilancia_alimentar.php?conteudo=pnsn

Instituto Nacional de Câncer. Estimate/2016 – Cancer Incidence in Brazil [Internet]. Rio de Janeiro: INCA; 2016. Available from: http://www.inca.gov.br/estimativa/2016/estimativa-2016-v11.pdf

Instituto Nacional de Câncer José Alencar Gomes da Silva. Sistema de informação do câncer: manual preliminar para apoio à implantação [Internet]. Rio de Janeiro, Brasil: INCA; 2013. Available from: http://www.fosp.saude.sp.gov.br:443/docs/Siscam/siscan_%20manual_preliminar.pdf

Instituto Nacional de Câncer José Alencar Gomes da Silva. Aumento de impostos sobre produtos de tabaco : apresentação e orientações técnicas [Internet]. Rio de Janeiro, Brasil: INCA; 2014. Available from: http://www1.inca.gov.br/tabagismo/31maio2014/manual-dia-mundial-sem-tabaco-2014.pdf

Instituto Nacional de Câncer José Alencar Gomes da Silva. OBSERVATÓRIO DA POLÍTICA NACIONAL DE CONTROLE DO TABACO [Internet]. 2014 [cited 2015 Feb 23]. Available from: http://www2.inca.gov.br/wps/wcm/connect/observatorio_controle_tabaco/site/status_politica/tratamento_tabagismo

Instituto Nacional de Câncer José Alencar Gomes da Silva, Coordenação de Prevenção e Vigilância. Estimativa 2014: Incidência de Câncer no Brasil [Internet]. Rio de Janeiro, Brasil; 2014. Available from: http://www.inca.gov.br/estimativa/2014/estimativa-24042014.pdf

J. Ferlay, I. Soerjomataram, M. Ervik. GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet]. Lyon, France: International Agency for Research on Cancer (IARC); 2012. Available from: http://globocan.iarc.fr/Default.aspx

Jacob Kligerman. Estimativas sobre a Incidência e Mortalidade por Câncer no Brasil - 2002. Rev. Bras. Cancerol. 2002;48(5):175–9.

Jaro MA. Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. J. Am. Stat. Assoc. - J AMER Stat. ASSN. 1989;84(406):414–20.

Jaro MA. Probabilistic linkage of large public health data files. Stat. Med. 1995;14(5–7):491–498.

Jemal A, Center MM, DeSantis C, Ward EM. Global patterns of cancer incidence and mortality rates and trends. Cancer Epidemiol. Biomarkers Prev. 2010;19(8):1893–907.

Jin L, Li C, Mehrotra S. Efficient record linkage in large data sets. Eighth Int. Conf. Database Syst. Adv. Appl. 2003 DASFAA 2003 Proc. 2003. p. 137–46.

Jr C, De KR, Coeli CM. Reclink: an application for database linkage implementing the probabilistic record linkage method. Cad. Saúde Pública. 2000 Jun;16(2):439–47.

Jurafsky D, Christopher Manning. Coursera - Natural Language Processing [Internet]. Coursera. [cited 2015 Feb 14]. Available from: https://www.coursera.org/lecture

Jurczyk P, Lu JJ, Xiong L, Cragan JD, Correa A. FRIL: A Tool for Comparative Record Linkage. AMIA. Annu. Symp. Proc. 2008;2008:440–4.

Kondrak G. N-gram similarity and distance. String Process. Inf. Retr. Springer; 2005. p. 115–26.

Leicester Gill. Methods for automatic record matching and linkage and their use in national statistics / Leicester Gill. - Version details - Trove [Internet]. 2001 [cited 2015 Jan 25]. Available from: http://trove.nla.gov.au/work/33417344?selectedversion=NBD23513617

Lenildo de Moura. Doença renal crônica terminal - uma proposta de monitoramento no Brasil. [Porto Alegre]: UFRGS; 2012.

Lenir Santos. SAÚDE: CONCEITO E AS ATRIBUIÇÕES DO SISTEMA ÚNICO DE SAÚDE [Internet]. Minist. Público - RS - Direitos Hum. 2005 [cited 2015 Feb 23]. Available from: http://www.mprs.mp.br/dirhum/doutrina/id387.htm

Levenshtein VI. Binary codes capable of correcting deletions, insertions, and reversals. Sov. Phys. Dokl. [Internet]. 1966 [cited 2015 Feb 13]. p. 707–710. Available from: https://gitlab.doc.ic.ac.uk/wm613/inidividual-

project/raw/4fbe43f863b229b50ca13bf5f59eb029ad71f6b6/reading/litreview/levenshtein66.pdf

Malta DC, França E, Abreu DMX, Perillo RD, Salmen MC, Teixeira RA, et al. Mortality due to noncommunicable diseases in Brazil, 1990 to 2015, according to estimates from the Global Burden of Disease study. Sao Paulo Med. J. 2017 Jun;135(3):213–21.

Malta DC, Jr S, Da JB. O Plano de Ações Estratégicas para o Enfrentamento das Doenças Crônicas Não Transmissíveis no Brasil e a definição das metas globais para o enfrentamento dessas doenças até 2025: uma revisão. Epidemiol. E Serviços Saúde. 2013 Mar;22(1):151–64.

Malta DC, Moura L de, Prado RR do, Escalante JC, Schmidt MI, Duncan BB. Mortalidade por doenças crônicas não transmissíveis no Brasil e suas regiões, 2000 a 2011. Epidemiol. E Serviços Saúde. 2014 Dec;23(4):599–608.

Malta DC, Moura L de, Prado RR do, Escalante JC, Schmidt MI, Duncan BB. Mortalidade por doenças crônicas não transmissíveis no Brasil e suas regiões, 2000 a 2011. Epidemiol. E Serviços Saúde. 2014 Dec;23(4):599–608.

Mendes A da CG, Lima MM de, Sá DA de, Oliveira LC de S, Maia LT de S. The use of the interrelation of data bases to improve information on child mortality in municipalities in the Brazilian State of Pernambuco. Rev. Bras. Saúde Materno Infant. 2012 Sep;12(3):243–9.

Ministério da Saúde. Plano de Ações Estratégicas para o Enfrentamento das Doenças Crônicas não Transmissíveis (DCNT) no Brasil 2011-2022 [Internet]. 2010 [cited 2012 Oct 12]. Available from: http://portal.saude.gov.br/portal/arquivos/pdf/cartilha_plano.pdf

Ministério da Saúde. Lei Antifumo [Internet]. 2014 [cited 2015 Feb 24]. Available from: http://portalarquivos.saude.gov.br/campanhas/leiantifumo/index.html

Ministério da Saúde Brasil. Portaria n º 2.043 de 11 de outubro de 1996. Implantação da Autorização de Procedimentos Ambulatoriais de Alta Complexidade/Custo (APAC) [Internet]. Diário Oficial da República Federativa do Brasil; 1996. Available from: sna.saude.gov.br/legisla/legisla/alta_cg/GM_2043_96alta_cg.doc

Ministério da Saúde Brasil. PORTARIA Nº 874, DE 16 DE MAIO DE 2013 - Ministério da Saúde Gabinete do Ministro [Internet]. 2013 [cited 2015 Feb 22]. Available from: http://bvsms.saude.gov.br/bvs/saudelegis/gm/2013/prt0874_16_05_2013.html

Ministério da Saúde Brasil. Portal da Saúde – Tabagismo – Ministério da Saúde [Internet]. 2013 [cited 2015 Feb 24]. Available from: http://portalsaude.saude.gov.br/index.php?option=com_content&view=article&id=6716&Itemid=290

Ministério da Saúde Brasil. PORTARIA Nº 571, DE 5 DE ABRIL DE 2013 - Ministério da Saúde - Gabinete do Ministro [Internet]. 2013 [cited 2015 Feb 23]. Available from: http://bvsms.saude.gov.br/bvs/saudelegis/gm/2013/prt0571_05_04_2013.html

Ministério da Saúde Brasil. MANUAL DE BASES TÉCNICAS DA ONCOLOGIA – SIA/SUS - SISTEMA DE INFORMAÇÕES AMBULATORIAIS [Internet]. Brasília, DF, Brasil; 2013. Available from: http://bvsms.saude.gov.br/bvs/publicacoes/inca/manual_oncologia_14edicao.pdf

Ministério da Saúde Brasil. PORTARIA Nº 7, DE 22 DE ABRIL DE 2014 - Ministério da Saúde [Internet]. 2014 [cited 2015 Feb 23]. Available from: http://bvsms.saude.gov.br/bvs/saudelegis/sctie/2014/prt0007_22_04_2014.html

Ministério da Saúde Brasil Brasil, Portal Brasil. SUS incorpora exame para pacientes com câncer — Portal Brasil [Internet]. SUS Inc. Exame Para Pacientes Com Câncer. 2014 [cited 2015 Feb 23]. Available from: http://www.brasil.gov.br/saude/2014/04/sus-incorpora-exame-para-pacientes-com-cancer

Ministério da Saúde Brasil, Fundação Nacional de Saúde Brasil. Manual de procedimento do sistema de informaçõıes sobre mortalidade [Internet]. Brasília; 2001. Available from: http://bvsms.saude.gov.br/bvs/publicacoes/sis_mortalidade.pdf

Ministério da Saúde Brasil, Gabinete do Ministro. PORTARIA Nº 876, DE 16 DE MAIO DE 2013 - Ministério da Saúde [Internet]. 2013 [cited 2015 Feb 23]. Available from: http://bvsms.saude.gov.br/bvs/saudelegis/gm/2013/prt0876_16_05_2013.html

Ministério da Saúde Brasil, Secretaria de Ciência, Tecnologia e Insumos Estratégicos. PORTARIA Nº 8, DE 14 DE ABRIL DE 2014 - Ministério da Saúde [Internet]. 2014 [cited 2015 Feb 23]. Available from: http://bvsms.saude.gov.br/bvs/saudelegis/sctie/2014/prt0008_14_04_2014.html

Ministério da Saúde Brasil, Secretaria de Ciência, Tecnologia e Insumos Estratégicos. PORTARIA Nº 9, DE 22 DE ABRIL DE 2014 - Ministério da Saúde [Internet]. 2014 [cited 2015 Feb 23]. Available from:

http://bvsms.saude.gov.br/bvs/saudelegis/sctie/2014/prt0009_22_04_2014.html

MINISTÉRIO DA SAÚDE BRASIL, Secretaria de Vigilância em Saúde, Departamento de Vigilância de Doenças e Agravos não Transmissíveis e Promoção da Saúde. Vigitel Brasil 2013 : vigilância de fatores de risco e proteção para doenças crônicas por inquérito telefônico [Internet]. Brasília, DF, Brasil; 2014 [cited 2015 Jan 22]. Available from: http://portalsaude.saude.gov.br/images/pdf/2014/dezembro/09/Vigitel-2013.pdf

MINISTÉRIO DA SAÚDE BRASIL, Secretaria de Vigilância em Saúde, Departamento de Vigilância de Doenças e Agravos não Transmissíveis e Promoção da Saúde. Vigitel Brasil 2016 : vigilância de fatores de risco e proteção para doenças crônicas por inquérito telefônico [Internet]. Brasília, DF, Brasil; 2017 [cited 2017 Jun 23]. Available from: http://portalsaude.saude.gov.br/images/pdf/2014/dezembro/09/Vigitel-2013.pdf

Ministério da Saúde BRASIL, Secretaria Executiva DATASUS, GEDINF – Gerência de Disseminação de Informações em Saúde. Disseminação de Informações do Sistema de Informações Ambulatoriais do SUS (SIASUS). Brasília, DF, Brasil; 2010.

Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Análise de Situação de Saúde. Brasil. SIM-Sistema de Informações de Mortalidade [Internet]. SIM-Sist. Informações Mortalidade. [cited 2015 Feb 6]. Available from: http://www2.datasus.gov.br/DATASUS/index.php?area=060701

MINISTÉRIO DA SAÚDE, SVS, Brasil. Portal da Saúde - Sistema de Informação sobre Mortalidade [Internet]. [cited 2015 Jan 27]. Available from: http://svs.aids.gov.br/cgiae/sim/

Moura L de, Curado MP, Simões EJ, Cezário AC, Urdaneta M. Avaliação do registro de câncer de base populacional do Município de Goiânia, Estado de Goiás, Brasil. Epidemiol. E Serviços Saúde. 2006 Dec;15(4):07-17.

Moura L de, Prestes IV, Duncan BB, Schmidt MI. Construção de base de dados nacional de pacientes em tratamento dialítico no Sistema Único de Saúde, 2000-2012. Epidemiol. E Serviços Saúde. 2014 Jun;23(2):227–38.

Notícias B da S Blog, Saúde, Ministério. Ministério da Saúde expande acesso ao tratamento oncológico [Internet]. Blog Saúde. 2014 [cited 2017 Jul 6]. Available from: http://www.blog.saude.gov.br/ta2rov

Oliveira GP de, Pinheiro RS, Coeli CM, Barreira D, Codenotti SB. Mortality information system for identifying underreported cases of tuberculosis in Brazil. Rev. Bras. Epidemiol. 2012 Sep;15(3):468–77.

Oliveira PPV de, Silva GA, Curado MP, Malta DC, Moura L de. Confiabilidade da causa basica de obito por cancer entre Sistema de Informacoes sobre Mortalidade do Brasil e Registro de Cancer de Base Populacional de Goiania, Goias, Brasil. Cad Saúde Pública. 2014;30(2):296–304.

Organisation mondiale de la santé, Alwan AAS, Agis T. Global status report on noncommunicable diseases 2010. Geneva: World Health Organization; 2011.

Organização Mundial de Saúde. WHO | UN Interagency Task Force on the Prevention and Control of Noncommunicable Diseases [Internet]. WHO. 2014 [cited 2015 Feb 24]. Available from: http://www.who.int/nmh/ncd-task-force/en/

Pita R, Pinto C, Melo P, Silva M, Barreto M, Rasella D. A Spark-based Workflow for Probabilistic Record Linkage of Healthcare Data. EDBTICDT Workshop. 2015. p. 17–26.

Portal Saúde, MS. Ministro da Saúde anuncia 11 prioridades para a oncologia [Internet]. Portal Saúde – Minist. Saúde – Wwwsaudegovbr. 2016 [cited 2017 Jul 6]. Available from: http://portalsaude.saude.gov.br/index.php/cidadao/principal/agencia-saude/25090-ministro-da-saude-anuncia-11-prioridades-para-a-oncologia

Presidência da República, Casa Civil, Subchefia para Assuntos Jurídicos. LEI Nº 12.546, DE 14 DE DEZEMBRO DE 2011. [Internet]. Planalto Gov. 2011 [cited 2015 Feb 24]. Available from: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12546.htm

Presidência da República, Casa Civil, Subchefia para Assuntos Jurídicos. DECRETO Nº 8.262, DE 31 DE MAIO DE 2014 [Internet]. 2014 [cited 2015 Feb 24]. Available from: http://www.planalto.gov.br/ccivil_03/_Ato2011-2014/2014/Decreto/D8262.htm

Randall SM, Ferrante AM, Boyd JH, Semmens JB. The effect of data cleaning on record linkage quality. BMC Med. Inform. Decis. Mak. 2013 Jun 5;13(1):64.

Rasella D, Aquino R, Santos CA, Paes-Sousa R, Barreto ML. Effect of a conditional cash transfer programme on childhood mortality: a

nationwide analysis of Brazilian municipalities. The Lancet. 2013 Jul 6;382(9886):57–64.

Ristad ES, Yianilos PN. Learning string-edit distance. IEEE Trans. Pattern Anal. Mach. Intell. 1998 May;20(5):522–32.

Sariyar M, Borg A. The RecordLinkage package: Detecting errors in data. R J. 2010;2(2):61–67.

Scatena JHG, Tanaka OY. Utilização do Sistema de Informações Hospitalares (SIH-SUS) e do Sistema de Informações Ambulatoriais (SIA-SUS) na análise da descentralização da saúde em Mato Grosso. Inf. Epidemiológico Sus. 2001 Mar;10(1):19–30.

Schnell R, Bachteler T, Reiher J. Privacy-preserving record linkage using Bloom filters. BMC Med. Inform. Decis. Mak. 2009 Aug 25;9(1):41.

Secretaria de Estado de Saúde do Distrito Federal. Sistema de Informações do Câncer - SISCAN [Internet]. 2014 [cited 2015 Feb 23]. Available from: http://www.saude.df.gov.br/outros-links/sistema-de-informacoes-do-cancer-siscan.html

Silva JFS da, Mattos IE. Avaliação da assistência oncológica de alta complexidade em um município da fronteira em Mato Grosso do Sul: uma proposta de cálculo de estimativas de cobertura. Cad Saúde Colet. 2012;20(3):314–20.

Stevens A, Schmidt MI, Duncan BB. Information–processing methods for mortality surveillance in the presence of varying levels of completeness and ill–defined codes of causes of death – the case of Brazil. Popul. Health Metr. 2014 Sep 26;12(1):24.

Stewart BW, Wild C, International Agency for Research on Cancer, World Health Organization, editors. World cancer report 2014. Lyon, France: International Agency for Research on Cancer; 2014.

Teixeira CL dos, Bloch KV, Klein CH, Coeli CM. Método de relacionamento de bancos de dados do Sistema de Informações sobre Mortalidade (SIM) e das autorizações de internação hospitalar (BDAIH) no Sistema Único de Saúde (SUS), na investigação de óbitos de causa mal-definida no Estado do Rio de Janeiro, Brasil, 1998. Epidemiol Serv Saúde. 2006;15(1):47–57.

Torra V, Domingo-Ferrer J. Record linkage methods for multidatabase data mining. Inf. Fusion Data Min. Springer; 2003. p. 101–32.

Torrens AW. Efetividade do Programa Bolsa Família na cura da tuberculose. 2015 Nov 10 [cited 2017 Jul 31]; Available from: http://repositorio.unb.br/handle/10482/18702

Torrens AW, Rasella D, Boccia D, Maciel ELN, Nery JS, Olson ZD, et al. Effectiveness of a conditional cash transfer programme on TB cure rate: a retrospective cohort study in Brazil. Trans. R. Soc. Trop. Med. Hyg. 2016 Mar 1;110(3):199–206.

Trading Economics - World Bank. Population - total in Low and middle income [Internet]. Low Middle Income - Popul. Total. 2017 [cited 2017 Jul 26]. Available from: https://tradingeconomics.com/low-and-middle-income/population-total-wb-data.html

TRIBUNAL DE CONTAS DA UNIÃO. Relatório de Auditoria – Política Nacional de Atenção Oncológica [Internet]. TRIBUNAL DE CONTAS DA UNIÃO; 2011. Available from: http://portal2.tcu.gov.br/portal/page/portal/TCU/comunidades/programas_governo/areas_atuacao/saude/Oncologia%20-%20relat%C3%B3rio%20-%20vers%C3%A3o%20final.pdf

Verykios VS, Moustakides GV, Elfeky MG. A Bayesian decision model for cost optimal record matching. VLDB J. 2003 May 1;12(1):28–40.

Wagner RA, Fischer MJ. The String-to-String Correction Problem. J. ACM. 1974 Jan 1;21(1):168–73.

William E. Winkler YT. An Application Of The Fellegi-Sunter Model Of Record Linkage To The 1990 U.S. Decennial Census. 1997;

Winkler WE. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. Proc. Sect. Surv. Res. Wachington, DC; 1990. p. 354–359.

Winkler WE. The State of Record Linkage and Current Research Problems. Statistical Research Division, U.S. Census Bureau; 1999.

Winkler WE. Overview of record linkage and current research directions. BUREAU OF THE CENSUS; 2006.

World Bank. Low & middle income | Data [Internet]. Low Middle Income Data. 2016 [cited 2017 Jul 26]. Available from: http://data.worldbank.org/income-level/low-and-middle-income

World Health Organization. Noncommunicable diseases country profiles 2011 [Internet]. Geneva, Switzerland: World Health Organization; 2011 [cited 2012 Oct 14]. Available from: http://www.who.int/nmh/publications/ncd_profiles_report.pdf

World Health Organization (WHO). A COMPREHENSIVE GLOBAL MONITORING FRAMEWORK INCLUDING INDICATORS AND A SET OF VOLUNTARY GLOBAL TARGETS FOR THE PREVENTION AND CONTROL OF NONCOMMUNICABALE DISEASES [Internet]. 2011

[cited 2012 Oct 12]. Available from: http://www.searo.who.int/LinkFiles/mhnd_GMF.pdf

WHO | 2008-2013 Action plan for the global strategy for the prevention and control of noncommunicable diseases [Internet]. 2009 [cited 2012 Oct 14]. Available from: http://www.who.int/nmh/publications/9789241597418/en/

WHO | Global Health Observatory Data Repository [Internet]. 2011 [cited 2012 Oct 14]. Available from: http://apps.who.int/ghodata/

FRIL: Fine-grained record linkage software [Internet]. 2011 [cited 2015 Jan 25]. Available from: http://fril.sourceforge.net/

Handbook of data quality: research and practice. New York: Springer; 2013.

MINISTÉRIO DA SAÚDE - Manual_Operacional_APAC_v_1_1.pdf [Internet]. [cited 2015 Jan 27]. Available from: ftp://arpoador.datasus.gov.br/siasus/Documentos/APAC/Manual_Operacional_APAC_v_1_1.pdf

MINISTÉRIO DA SAÚDE - Manual_Operacional_BPA.pdf [Internet]. [cited 2015 Jan 27]. Available from: ftp://arpoador.datasus.gov.br/siasus/Documentos/BPA/Manual_Operacional_BPA.pdf

MINISTÉRIO DA SAÚDE - Manual_Operacional_SIA_v_1.pdf [Internet]. [cited 2015 Jan 27]. Available from: ftp://arpoador.datasus.gov.br/siasus/Documentos/sia/Manual_Operacional_SIA_v_1.pdf

Practical Cryptography [Internet]. [cited 2015 Jan 25]. Available from: http://practicalcryptography.com/cryptanalysis/text-characterisation/monogram-bigram-and-trigram-frequency-counts/

Cryptanalysis of basic Bloom Filters used for Privacy Preserving Record Linkage - -download=wp-grlc-2014-04.pdf [Internet]. [cited 2015 Jan 25]. Available from: http://www.record-linkage.de/-download=wp-grlc-2014-04.pdf

Dedupe Software / Record Linkage Software by The Link King FREE ! [Internet]. [cited 2015 Jan 25]. Available from: http://www.the-link-king.com/

BigData.pdf [Internet]. [cited 2015 Jan 25]. Available from: http://cran.r-project.org/web/packages/RecordLinkage/vignettes/BigData.pdf

ARTIGO 1

Pareamento de registros das grandes bases de dados do Sistema Único de Saúde para possibilitar análise longitudinal de pacientes com câncer em tratamento por quimio- ou radioterapia, de 2000 a 2012

Record linkage of large databases of the Brazilian National Health System for longitudinal analyses of cancer patients undergoing chemo- or radiotherapy, 2000 to 2012

Isaías Valente Prestes, Doutorando em Epidemiologia pela UFRGS.

Bruce Bartholow Duncan, Federal University of Rio Grande do Sul, Postgraduate Program in Epidemiology

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL (UFRGS)

A ser enviado ao Population Health Metrics

TITLE

Record linkage of large databases of the Brazilian National Health System for longitudinal analyses of cancer patients undergoing chemo- or radiotherapy, 2000 to 2012

Abstract

Background – Investigation of cancer prognosis and treatment in Brazil could be advanced by greater use of administrative data collected by the national health system (SUS). Yet to date, these data have not been organized at the individual patient level to permit epidemiologic analyses. The objective of this study is to describe the construction of a cohort of patients whose treatment for cancer was financed by the SUS, utilizing data from the High Complexity Procedures Authorization (APAC) database, an administrative database of the Ministry of Health.

Methods: We applied deterministic and probabilistic record linkage, using the Statistical Analysis System (SAS) and the Fine-Grained Record Integration and Linkage Tool (FRIL), to organize multiple separate APAC records into a format permitting follow-up of cancer patients over the period of the study, from 2000 to 2012. We then added initial treatment information and patient survival through linking records of these cancer patients available within the national public hospitalization database (AIH) and the national mortality information system (SIM).

Results: We linked 9,170 APAC files, identifying 23,048,694 records referring to chemo- or radiotherapy. Among these, we recovered 3,168 (6.26%) of the 52,808 records (0.23%) lacking key variables, or with invalid values for these identifiers. From 2001-2012, approximately 1.5 million individuals initiated cancer chemo- or radiotherapy financed by the SUS. The record linkage between the APAC and SIM databases recovered mortality information of patients related to malignant neoplasm between 2001 and 2013.

Conclusion: The methodology used permitted the construction of a database capable of monitoring the flow of cancer patients whose treatment was financed by the SUS. This approach can also be applied for monitoring procedures for other diseases in Brazil or in countries where similar, difficult to analyze, administrative databases have been constructed but never adequately analyzed.

Keywords: Cancer; Health Information Systems; Epidemiological Surveillance; Record Linkage.

Background

Cancer is the second leading cause of morbidity and mortality worldwide, accounting for 8.8 million deaths in 2015[1]. Approximately 70% of deaths from cancer occur in low- and middle-income countries. Yet the treatment of cancer in these countries has received relatively little attention, in part due to the lack of available data sources for investigation.

In Brazil, the Brazilian National Cancer Institute (INCA) estimated that a total of 596,000 new cases of cancer occurred nationwide in 2016, 295,200 among men and 300,800 among women[2,3].According to information from the 2013 Brazilian National Health Survey, only 27.9% of Brazilians have a private health plan, the remaining 72.1% receive care exclusively through the Brazilian national health system (Sistema Único de Saúde, or SUS). The SUS increased spending on cancer treatments by 66% over the 2010 and 2015 period - from US$ 650.4 million to US$ 1.08 billion, using the exchange rate of July 1, 2017. This allowed increased coverage for procedures such as oncologic surgeries, chemotherapy, radiotherapy, hormone therapy and palliative care [4].

Taking into account the burden of cancer and the cost of its treatment to the SUS, carefully designed monitoring of data on public treatment of cancer is essential to evaluate the implementation and progress of these actions [5,6]. However, access to information sources, essential for effective monitoring and evaluation, has been very limited to date. This problem can be minimized by converting administrative databases – useful sources of information for cancer incidence monitoring [7] – into epidemiologically useful databases. The Brazilian public health system registers all expenses related to publicly-financed cancer treatment in the High Complexity Procedure Authorization/Radio and Chemotherapy subsystem (Autorização de Procedimentos de Alta Complexidade, series AR and AQ) of the Outpatient Information System (Sistema de Informação Ambulatorial, or SIA)[8], except for those expenses of cancer related to surgery and some other specific in-hospital procedures. Data on these are recorded in the Hospitalization Information System (Sistema de Informações Hospitalares or SIH/SUS), an

administrative database which records all medical care originating from hospitalizations financed by the SUS, permitting payments to health facilities. However, due of a non-unique patient identifier number over time (2000-2012), these data have been unavailable for epidemiological monitoring (e.g. incidence, prevalence, and mortality) and evaluation of cancer treatment. We developed a methodology to create a unified database by linking patient records within the APAC cancer subsystem[9] and between APAC, on the one hand, and the SIH and SIM systems, on the other, so as to permit such analyses.

The objective of this study is to describe the creation, though record linkage, of this national database of cancer patients whose care was financed by the SUS.

Methods

Oncological treatment data sources

We obtained High Complexity Procedures Authorization (SIA / APAC) databases with nominal data from the Brazilian Ministry of Health. These databases, produced by DATASUS, the Ministry´s data processing unit, cover the period from 2000 to 2012, and include chemo- and radiotherapy.

We combined this chemo- and radiotherapy information into a single cancer database, merging 9,170 period and Brazilian state specific APAC files (Figure 1). APAC was initially structured and used from 2000 to 2007 in the format of three separate databases (APAC main form [APA], Chemotherapy APAC data [PQ], Radiotherapy APAC data [PR]) for a given patient reimbursement, each covering a different aspect of his registration into the system or the care he received. In 2008, these three separate databases were merged into two: Chemotherapy APAC data (AQ) and Radiotherapy APAC data (AR). Each APAC record covers a period of at most 3 months. Thus, a given patient often accumulates dozens of separate records over his period of treatment. Patient identifiers used in the different records varied over the period, initially being the patient´s taxpayer identification number (Cadastro de Pessoas Físicas, or CPF), and

later the Brazilian national patient identification number (Cartão Nacional de Saúde, or CNS). Additionally, distribution of CNSs was such that it was not uncommon for a given individual to have more than one CNS.

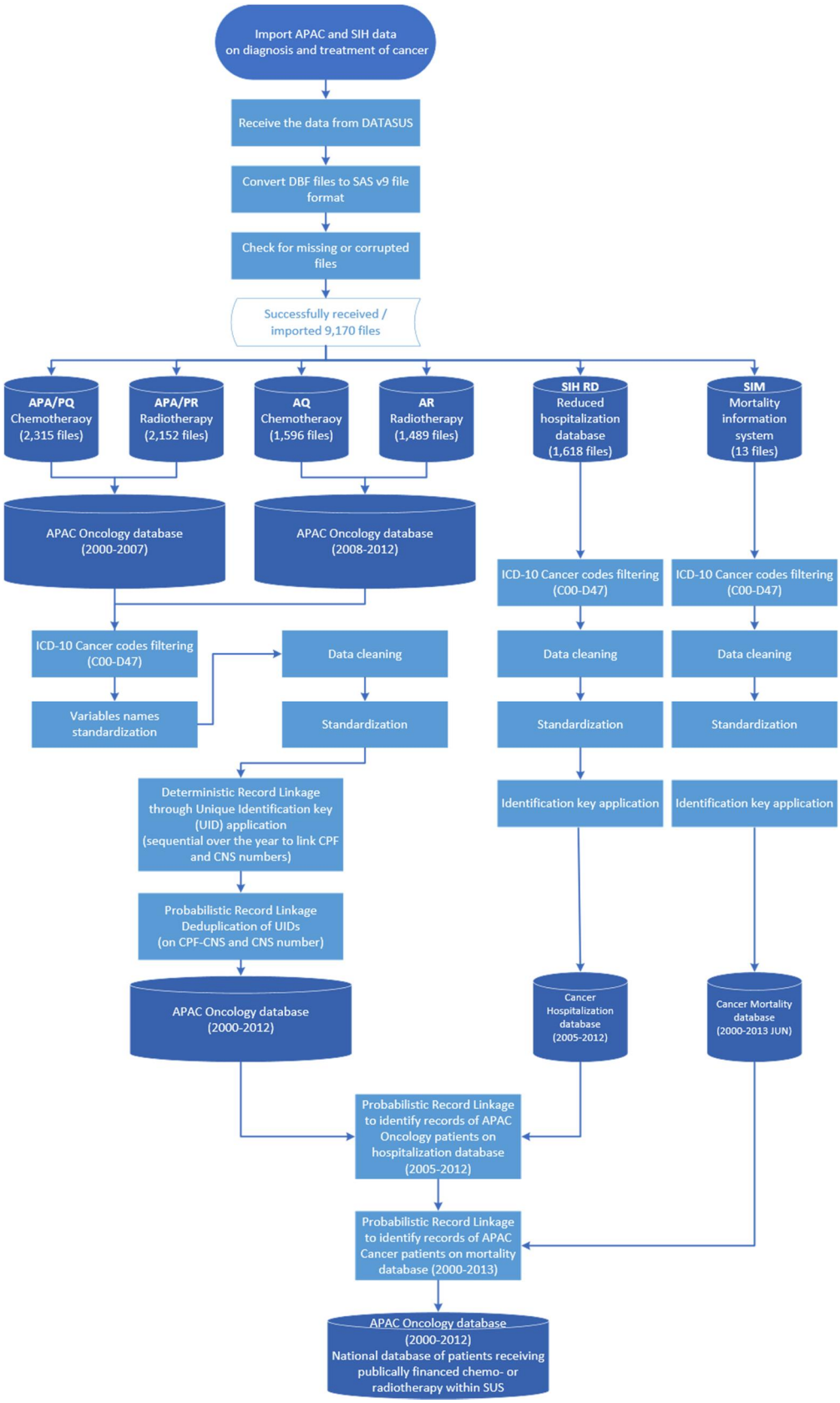


FIGURE 1 – APAC Oncology database creation, indicating all record linkage process carried out.

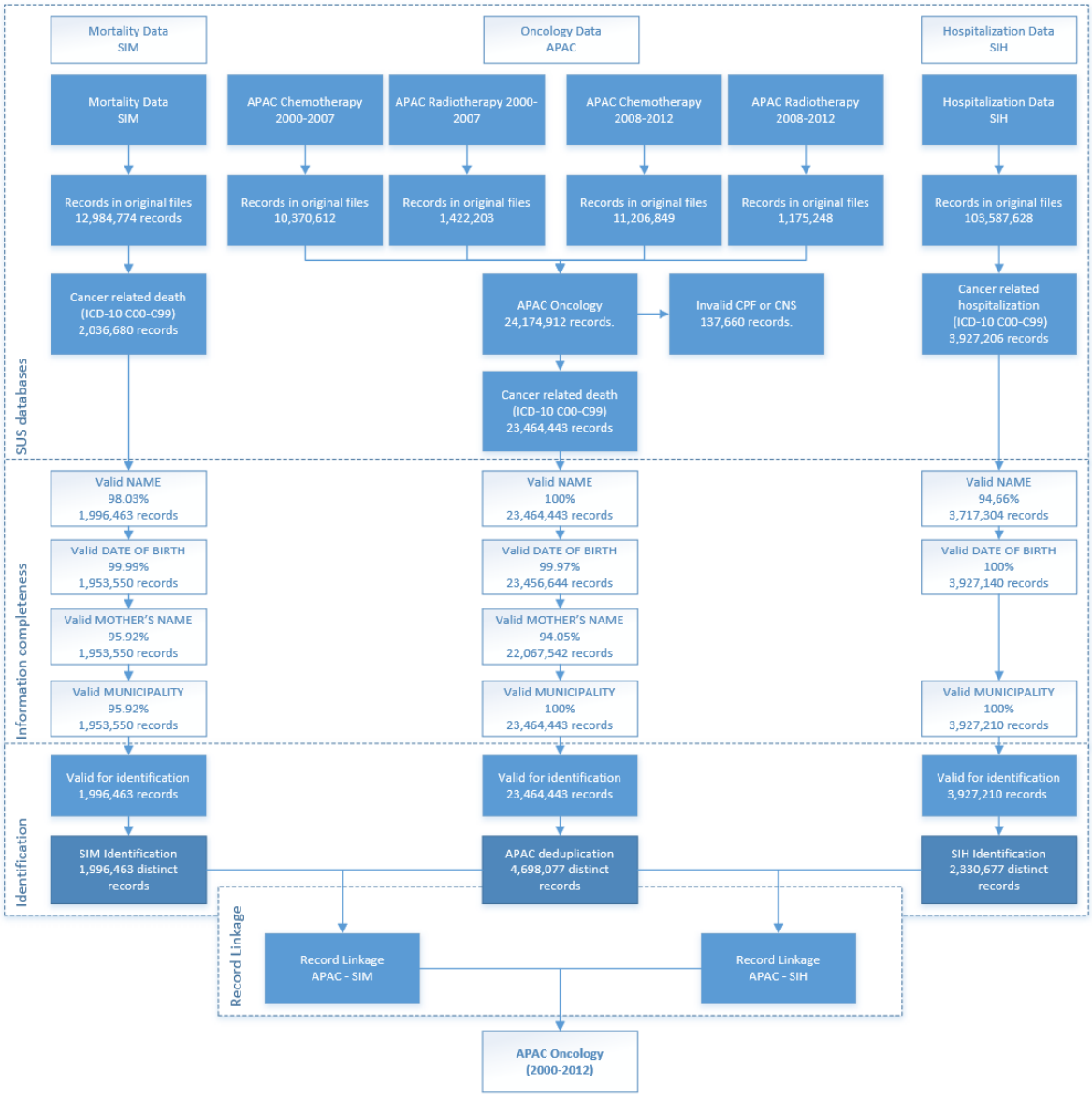


FIGURE 2 – Information sources, record inclusion criteria and merging of data sources on chemo- and radiotherapy (APAC), hospitalizations (AIH), and mortality (SIM) of cancer patients whose treatment was financed by SUS.

Each APAC data record concentrates information of chemo- or radiotherapy performed to treat a specific cancer in a SUS-associated hospital or clinic. However, as APAC can be used for a maximum of 3 months, additional APAC records are generated when the treatment duration covers a longer period.

We initially detected 247 missing files for APAC Chemotherapy – 223 during the period 2000-2007 and 24 during the period 2008-2012 – and 400 missing files for APAC Radiotherapy – 371 over 2000-2007 and 29 over 2008-2012. An investigation revealed two reasons for this: 1) no SUS-financed health care establishment provided cancer care over a given month in a specific Brazilian state; and 2) a delay in processing the APAC records occurred, with all records for the missing period being reported in the next month´s file. Thus, these missing files were of no import for our analyses.

We merged data on chemo- (APAC series PQ, AQ) and radiotherapy (APAC series PR and AR) into a single SAS dataset, which we denominated APAC Oncology. As the APAC system is also used to record reimbursable actions related to other diseases, we then selected records classified as related to cancer (Figure 2) using codes C00 to D47 of the 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD-10). We examined frequency tables to evaluate quality of data values (completeness of the identifying information), and to detect possible invalid or inconsistent information.

After data exploration, we carried out further data cleaning and standardization processes following recommended procedures [10–12]. Data cleaning involved data record deduplication and subsequent manual inspection to classify removed records. We standardized all string data converting text field characters in uppercase, removing special characters (including invisible hexadecimal codes), and parsing text strings (e.g. first name, last name, middle name, etc.). For variables containing names (patient and patient's mother's names), we removed common prepositions inside names – e.g. "José Genuíno da Silva" → "JOSE GENUINO

SILVA). We also standardized the database representation (as the software shows a given variable values to the user) of each variable, and adjusted numerical valid ranges or finite valid values, including the conversion of all dates to the format YYYYMMDD. The APAC dates were initially imported as DDMMYYYY, and the SIM dates as YYYYMMDD)[10]. We created a special subset of records lacking personal identifiers – CPF and CNS – for attempts to recuperate their identification using deterministic and probabilistic record linkage.

Over time, the APAC database structure has changed. As a result, the same information may be stored in variables with different names, e.g. date of birth was stored in AQ_DATANASC between 2000 and 2007, and in AP_DATANASC between 2008 through 2012. We created derived variables to aggregate these variables, permitting the compatibility of information across the time series.

Hospitalization data source

SIH records are summaries of the source document for hospitalization – the Authorization for Hospital Admittance (Autorização de Internação Hospitalar, or AIH). We compiled information of hospitalizations from SIH for records with valid information inserted from July 1st, 2004 through December 31th, 2012, creating an auxiliary database of hospital admissions having cancer as a cause for admission (Figure 2). To do so, we merged 1,618 files obtained from DATASUS, after filtering completeness and including only those records with the following validation criteria: hospitalization date between July 2004 and December 2012 (DT_INTER), and ICD-10 codes C00 to D47 for 1) the main diagnosis (DIAG-PRINC), 2) the secondary diagnosis (DIAG_SECUN), 3) the cause of death (CID_MORTE), or 4) an associated cause of death (CID_ASSO).

Mortality data source

The mortality database for this study resulted from the importation of the 13 annual SIM files covering the period from 2001 to July, 2013. These databases were provided by

DATASUS and the Secretaria de Vigilância Sanitária (SVS-MS). We selected mortality records for linkage having ICD-10 codes C00 to D47 on any of the cause of death lines. The SIM records with missing patient name, or invalid names (e.g. “White man”, “Ignored”, “Unknown”, “Not available”) were excluded.

Record linkage procedures

We performed deterministic and probabilistic record linkage [7,9,13,14], in order 1) to create the unique patient identification number (UID) for the APAC Oncology database, and to link identified patients with the relatively small subset of records without CPF or CNS numbers, 2) to associate information from hospitalizations (SIH) with the APAC Oncology database, and 3) to add information on deaths to the APAC Oncology database via linkage between APAC Oncology and the national mortality information system.

Prior to performing this linkage, we by performing a special exercise to identify parameter settings that optimized correct identification of pairs, as defined by ROC curve analysis, for our probabilistic linkage algorithm[15]. Additionally, we evaluated cut-off score values, above which all pairs are declared true matches in an automatic record linkage decision model, using two standard metrics: accuracy and the F-measure. The F-measure of a classification system is defined as the weighted harmonic mean of its precision (positive predictive value), and recall (sensitivity). This measure can be calculated through the expression F-measure = (2 * Precision) / (Precision + Recall). The accuracy measure reflects the percentage of correct predictions from all predictions made by a model, and it is defined by the expression Accuracy = (# true positives + # true negatives) / (total # of link pairs). For measuring improvements across these experiments, accuracy and the F-measure are recognized as being better than sensitivity and specificity. The utility of specificity in particular is limited by the large number of true negatives (non-matches) vis-a-vis the low frequency of matches [16]. In this process, we used the full naïve comparison method[17] to all tested combinations of weighting

parameters values. A full naïve algorithm compares each pair of records to find matches, and may produce more trustable results, although it is disadvantageous as it increases processing time[18].

We performed this exercise over all distinct combinations of patient record values within the 2004 national APAC oncology database. We examined combinations of weight values, these being the percentage of the final confidence score for a two-record link allocated to a given variable, ranging from 50% to 100% in increments of 5%, for the patient´s name, and from 0% to 50%, in increments of 10%, for patient´s mother's name, date of birth, and municipality of residence. Furthermore, we applied Levenshtein edit distance[19,20] to access similarity among record pairs for all match variables, with a special rejection threshold if similarity was less than 40% (scoring zero for this specific variable in the composition of the score).

In this investigation of the best weighting of parameters, the classification of true non-match and true match pairs considered *a priori* known and reliable information based on CPF and CNS numbers. Thus, we classified true match pairs of records where both CPF numbers, or both CNS numbers, were equal. A true non-match was defined when both CPF/CNS values of the record pair were not equal. This procedure allowed us to classify the record pairs, for a given decision model, in true positive, true negative, false positive, and false negative. In order to compare our findings with this approach to those with the most frequently used method to define record linkage pair score, we evaluated the best set of weights for record linkage using the Expectation–Maximization (EM) algorithm[16] - an iterative method to find maximum likelihood or maximum *a posteriori* values for variable weights. The algorithm learned, using 50% of data for 2004, applying the all-to-all search method.

We performed the process of unique identifier (UID) implementation, to identify clusters of patient records throughout a deterministic linkage, sequentially in three periods of years, established according to the set of valid patient identification numbers. The first period,

between 2000 through 2004, contains patient records having only CPF as identifier (primary key). The second period, between 2004 through 2007, is a transitional period of patient identifiers, meaning that patients may have CPF and / or CNS identifying his records. The third and last period, between 2008 through 2012, contains only CNS as a patient identifier, although a given patient may present multiple CNS numbers (not unique key).

For the first period (2000-2003), a UID number was associated with each valid CPF number present in the patients' records. Then, throughout the second period (2004-2007), we performed the association of the UIDs, already defined in the first period, with new records with CPF numbers -meaning, we applied the UID deterministically on prevalent patients, aiming to aggregate clusters of records defining the patient´s treatment history. The next step was to assign new UID numbers to remaining CPF numbers not assigned to a previous UID (incident cases) over this second period (2004-2007). At this point, based on patients' records with valid filled out CPF and CNF fields in the second period, we created a list of UID numbers with valid CPF and CNS values, linking identifiers. Using this linking information, we proceeded to a new step: we applied the UIDs with valid CPF and CNS values, from the second period, on the records of the third period (2008-2012) with a matching CNS number. Finally, we assign new UID numbers to the CNS numbers of patients not having been previously assigned a UID (incident cases) over the third period (2008-2012).

Next, applying the best variable weights and their optimum cut-offs determined as described above, we carried out an internal probabilistic record linkage on the APAC Oncology database, to correct duplications, that is to say, pseudo distinct clusters of registers generated by deterministic record linkage mentioned above. We performed these record linkages in temporal order, within a time horizon of 3 years, to reduce the probability of false positive links, e.g. comparison and possible linkage of APAC clusters of patients initiating treatment in 2004 to APAC clusters prevalent from 2004 through 2006, never clusters of 2000 through 2003, or

clusters of 2007 through 2012. A full naïve comparison method was applied without any other blocking method.

After complete record linkage on the APAC Oncology database to identify patients, we inspected frequency tables of distinct combinations of name, date of birth and mothers name within UIDs. UIDs with a large number of distinct elements received the classification of a suspicious cluster, were manually inspected, and were divided into different UIDs if deemed appropriated.

We next conducted a probabilistic record linkage of the APAC Oncology database with hospitalization information from the SIH in order to capture additional treatments done in hospitals, and to obtain additional information concerning the cancer treatment start date. This procedure used patient name, date of birth, and mother's name as matching variables for records pairing. We performed this record linkage in temporal order, to reduce the probability of false positive links to hospitalization and the number of comparisons performed, e.g. we linked APAC records of patients initiating treatment in 2008 only to SIH records from 2007 through 2008, never records of 2004 through 2006, nor of 2009 through 2012. A full naïve comparison method was also applied here without any other blocking method.

Finally, we performed probabilistic record linkage of APAC Oncology records with the mortality database (SIM) to correct the underreporting of deaths in the APAC data, in the process generating a new, more accurate derived variable for the presence of cancer-related deaths. The matching variables used for pairing here were patient's and patient´s mother's names, date of birth, and patient municipality of residence, applying the optimal weights and cut-off score point detected during the APAC linkage process. Processes of standardizing (cleaning) the data also followed the same pattern reported above. We performed these record linkages in temporal order, to reduce the probability of false positive links to deaths, e.g. linkage of APAC records of patients initiating treatment in 2004 to SIM data investigated only SIM

records from 2004 through 2013, never records of 2001 through 2003. A full naïve comparison method was also applied here without any other blocking method. For quality control purposes, we did not exclude patients known from APAC or SIH to have died, thus allowing use of this output to evaluate our ability to detect a death within these other systems.

Software

The APAC Oncology database was developed with and controlled by the Statistical Analysis System (SAS® 9.4). Record linkage was performed in FRIL (Fine-grained record linkage software) version 2.1.5[16]. The record linkage process was conducted on five HP Compaq 6300, Intel i7 3770 (3.4 GHz), 8Gb RAM, and one HP Workstation Z210, running Microsoft Windows 7 PRO 64 Bits. All hard disks, including external hard drives used as backup units, were encrypted units managed initially by TrueCrypt[21,22], this being replaced by VeraCrypt[23] after security issued arose concerning TrueCrypt.

Ethics

The Ethics in Research Committee of the Hospital de Clínicas de Porto Alegre approved the study on 10/03/2010 (No. 100056). As analyses are based on surveillance databases from the Ministry of Health, no patient consent was necessary. All those with access to the data signed agreements to maintain the confidentiality of the information.

Results

The APAC Oncology database contained 23,649,650 records with mention of cancer (C00-C99), 99.2% of which showing sufficient information to assign a UID. The record linkage performed organized these records into clusters with distinct UIDs for each of the 1,742,149 individuals who received cancer chemo- or radiotherapy financed by the SUS from 2000 to 2012. The APAC chemotherapy database (PQ, AQ series, Figure 1) contributed 89.3% of all records, and the radiotherapy database (PR, AR series) the remaining 10.7%. Although, 190,726 records

lacked the CPF and CNS key variables (or had an invalid value for these identifiers), 69% of these records were recovered after information processing.

The SUS, from 2000 through 2012, treated 743,574 (42.7%) patients with just chemotherapy, 420,238 (24.1%) patients with just radiotherapy, and 578,337 (33.2%) patients with combined chemo and radiotherapy.

The SIM database contributed 646,689 cancer-related records – a total of 15.3% of all reported deaths on SIM database between years 2001-2013. The SIH database contributed 3,927,206 records with cancer ICD-10 codes, representing 3.8% of all hospitalization records over the years 2005 to 2012.

Among records in the APAC Oncology database, 5.8% had missing information for mother's name and <0.2% had missing information for any of the other variables important for linkage of cancer patients´ records (Table 1): patient's name, date of birth, municipality of residence and primary cause of cancer, revealing an acceptable information completeness. However, secondary variables, such as date of diagnosis and date of initial treatment presented a varying degree of completeness and detail of information. Between 2000 through 2007, date of diagnosis was expressed in months, and was present and adequate in only 24.2% of records. However, from 2008 through 2012, this variable was recorded in days and was 99.9% complete. Similarly, between 2000 through 2007, the date of initial treatment was present (expressed in month only) and adequate in only 24% of records, but it´s completeness improved to 99.9% over the period of 2008 through 2012. Patients on second or further treatment showed a completeness of 14.4% for the date of their first treatment over the period of 2000 through 2007, and of 45.9% when this was recorded between 2008 through 2012.

Within the hospitalization and mortality data, identifier variables presented a maximum of 5.34% of missing information, and the main information related to the purpose of linkage was nearly always present (Table 1).

TABLE 1 – Linkage variable quality after data standardization (ICD-10 C00-C99)

APAC Oncology, Chemo-Radiotherapy, (2000-2012)			
	Valid values (%)	Invalid values (%)	Missing (%)
Name	100.00	0.00	0.00
Date of birth	99.97	0.01	0.03
Mother's name	94.05	5.85	0.10
Municipality of residence	100.00	0.00	0.00
Sex	100.00	0.00	0.00
Primary cause (ICD-10)	100.00	0.00	0.00
Date of diagnosis*	99.97	0.02	0.01
Date of initial treatment*	99.99	0.01	0.00
Stage	90.89	0.00	9.11
SIH - Hospitalization (2005-2012)			
	Valid values (%)	Invalid values (%)	Missing (%)
Name	94.66	5.34	0.00
Date of birth	100.00	0.00	0.00
Municipality of residence	100.00	0.00	0.00
Sex	100.00	0.00	0.00
Primary Cause (ICD-10)	100.00	0.00	0.00
Secondary Cause (ICD-10)	85.29	6.42	7.89
Date of hospitalization	100.00	0.00	0.00
Date of hospital discharge	100.00	0.00	0.00
SIM - Mortality (2001-2013 1/2)			
	Valid values (%)	Invalid values (%)	Missing (%)
Name	98.00	0.05	1.95
Date of birth	99.99	0.01	0.00
Mother's name	95.92	0.10	4.08
Municipality of residence	95.85	0.10	4.05
Sex	99.98	0.01	0.01
Date of death	100.00	0.00	0.00

* Became available, in days (format mm/dd/aaaa) in APAC database in 2008.

The set of weight combinations tested for the linkage variables (TABLE 2) with the APAC Oncology data revealed a curvilinear pattern, with only one maximum point (65.5%) for measures of quality. The optimal weighting – that which maximizes accuracy, the F-measure, and the AUC measures – is the assignment of 50% of the final score to the patient´s name, 20% to the patient's mother´s name, 20% to the date of birth, and 10% to the municipality of residence. Although the above weight setting, <50,20,20,10>, appeared to be the best solution based on point estimates of quality measures, confidence intervals of the weight settings <50,20,30,0> and <50,20,20,10> overlap for recall (sensitivity), specificity, and precision (positive predictive value). F-measure and accuracy measures differed by only 1% between the best setting and these other two settings.

The EM (expectation-maximization) algorithm suggested weights for the final score of 25% for the name of patient, 28% for the patient mother's name, 32% for date of birth, and 15% for the patient's municipality of residence. The applied empirical methodology, compared to the EM algorithm, resulted in an increase of 367,626 true negative pairs (6.673%), 4,491 true positive pairs (0.6%) and 2,070 false positive pairs (+ 125.6%). On the other hand, there was a reduction of 8,359 false negative pairs (-34.6%). Applying the EM weights to the record linkage, we found a recall (sensitivity) of 96.8% (96.76 – 96.84), accuracy of 0.966, precision (positive predictive value) of 99.77% (99.75 – 99.79), F-measure of 0.9826, AOC of 92.13 (91.96 – 92.31), and the best cut-off point at 71% of the confidence score. All quality measures of our empirical optimum setting dominated quality measures of the EM algorithm.

TABLE 2 – Accuracy and precision of different weight combinations for probabilistic record linkage quality as calculated for record pairs with scores of over 50% confidence for a true-match.

Record linkage weights				Cut-off Score	Sensitivity % (95% CI)	Specificity % (95% CI)	Predictive Positive Value % (95% CI)	Area under the ROC curve % (95% CI)	F-measure	Accuracy
Name	*Mother's name*	*DOB*	*Municipality of residence*							
50	0	0	50	68.5	95.42 (95.37-95.47)	98.71 (98.70-98.72)	84.35 (84.28-84.43)	98.78 (98.76-98.80)	0.895	0.985
50	0	10	40	77.5	93.92 (93.87-93.97)	98.06 (98.02-98.09)	98.12 (98.09-98.15)	98.79 (98.78-98.81)	0.960	0.959
50	0	20	30	68.5	97.49 (97.45-97.52)	96.65 (96.60-96.70)	97.82 (97.79-97.85)	99.27 (99.26-99.29)	0.977	0.972
50	0	30	20	70.5	97.15 (97.11-97.19)	99.07 (99.04-99.10)	99.49 (99.47-99.50)	99.41 (99.40-99.42)	0.983	0.978
50	0	40	10	67.5	97.58 (97.54-97.61)	98.81 (98.78-98.84)	99.35 (99.33-99.36)	99.42 (99.41-99.43)	0.985	0.980
50	0	50	0	71.5	96.74 (96.70-96.78)	99.02 (99.00-99.05)	99.34 (99.32-99.36)	99.31 (99.29-99.32)	0.980	0.976
50	10	0	40	76.5	94.03 (93.98-94.09)	97.91 (97.88-97.95)	97.96 (97.93-98.00)	98.67 (98.65-98.69)	0.960	0.959
50	10	10	30	68.5	96.81 (96.78-96.85)	96.73 (96.68-96.78)	97.88 (97.85-97.91)	99.37 (99.36-99.38)	0.973	0.968
50	10	20	20	71.5	96.71 (96.67-96.75)	99.45 (99.42-99.47)	99.70 (99.69-99.71)	99.54 (99.53-99.55)	0.982	0.977
50	10	40	0	67.5	97.54 (97.50-97.57)	98.91 (98.88-98.94)	99.41 (99.39-99.43)	99.52 (99.51-99.53)	0.985	0.980
50	20	0	30	68.5	96.74 (96.70-96.78)	96.71 (96.67-96.76)	97.77 (97.74-97.80)	99.15 (99.13-99.16)	0.973	0.967
50	20	10	20	70.5	96.50 (96.45-96.54)	99.33 (99.31-99.36)	99.64 (99.62-99.65)	99.50 (99.49-99.51)	0.980	0.975
50	20	20	10	65.5	97.88 (97.85-97.92)	99.01 (98.98-99.04)	99.50 (99.49-99.52)	99.56 (99.55-99.57)	0.987	0.983
50	20	30	0	65.5	97.87 (97.83-97.90)	98.83 (98.80-98.87)	99.40 (99.39-99.42)	99.53 (99.52-99.54)	0.986	0.982
50	30	0	20	68.5	96.67 (96.63-96.71)	97.29 (97.25-97.34)	98.44 (98.41-98.47)	99.25 (99.24-99.27)	0.975	0.969
50	30	10	10	65.5	97.31 (97.27-97.34)	98.01 (97.97-98.05)	98.98 (98.95-99.00)	99.49 (99.48-99.50)	0.981	0.975
50	30	20	0	68.5	96.94 (96.90-96.98)	98.82 (98.78-98.85)	99.39 (99.37-99.41)	99.48 (99.46-99.49)	0.981	0.976
50	40	0	10	59.5	98.08 (98.05-98.11)	92.48 (92.40-92.56)	95.91 (95.86-95.95)	99.02 (99.01-99.04)	0.970	0.961
50	40	10	0	59.5	98.21 (98.18-98.24)	93.76 (93.68-93.83)	96.79 (96.75-96.83)	99.18 (99.17-99.20)	0.975	0.967
50	50	0	0	68.5	91.54 (91.47-91.60)	97.22 (97.18-97.26)	97.62 (97.59-97.66)	97.13 (97.11-97.16)	0.945	0.941
60	0	0	40	83.5	93.49 (93.44-93.55)	98.05 (98.02-98.08)	97.47 (97.43-97.50)	97.83 (97.81-97.86)	0.954	0.960
60	0	10	30	69.5	97.41 (97.38-97.45)	96.09 (96.04-96.13)	96.36 (96.31-96.40)	98.84 (98.82-98.86)	0.969	0.968
60	0	20	20	73.5	97.05 (97.02-97.09)	98.19 (98.16-98.23)	98.54 (98.51-98.57)	98.92 (98.91-98.94)	0.978	0.976
60	0	30	10	72.5	97.05 (97.01-97.08)	99.32 (99.30-99.34)	99.47 (99.45-99.49)	98.94 (98.92-98.95)	0.982	0.980
60	0	40	0	77.5	96.09 (96.04-96.13)	99.49 (99.47-99.51)	99.58 (99.56-99.59)	98.91 (98.89-98.92)	0.978	0.976
60	10	0	30	69.5	96.76 (96.72-96.80)	96.00 (95.96-96.05)	96.28 (96.24-96.32)	98.67 (98.66-98.69)	0.965	0.964
60	10	10	20	69.5	97.49 (97.46-97.53)	97.51 (97.47-97.54)	98.03 (98.00-98.06)	98.90 (98.88-98.92)	0.978	0.975
60	10	20	10	71.5	97.21 (97.17-97.25)	99.35 (99.33-99.37)	99.51 (99.50-99.53)	98.92 (98.91-98.94)	0.983	0.981
60	10	30	0	72.5	96.97 (96.93-97.00)	99.35 (99.33-99.37)	99.50 (99.49-99.52)	98.92 (98.90-98.93)	0.982	0.980
60	20	0	20	73.5	96.21 (96.17-96.25)	98.03 (98.00-98.07)	98.41 (98.39-98.44)	98.71 (98.70-98.73)	0.973	0.970
60	20	10	10	72.5	96.58 (96.54-96.62)	99.13 (99.11-99.16)	99.35 (99.33-99.37)	98.87 (98.86-98.89)	0.979	0.977
60	20	20	0	72.5	96.91 (96.87-96.95)	99.06 (99.03-99.08)	99.30 (99.29-99.32)	98.86 (98.84-98.88)	0.981	0.978
60	30	0	10	69.5	96.75 (96.71-96.79)	96.19 (96.14-96.24)	97.17 (97.13-97.21)	98.55 (98.53-98.57)	0.970	0.965
60	30	10	0	69.5	96.94 (96.90-96.98)	97.28 (97.24-97.32)	98.00 (97.97-98.03)	98.68 (98.66-98.70)	0.975	0.971
60	40	0	0	69.5	92.18 (92.12-92.24)	96.28 (96.23-96.33)	97.07 (97.03-97.11)	97.08 (97.05-97.10)	0.946	0.939
70	0	0	30	79.5	94.02 (93.97-94.08)	98.20 (98.18-98.23)	97.20 (97.17-97.24)	97.88 (97.85-97.90)	0.956	0.965
70	0	10	20	74.5	97.02 (96.99-97.06)	98.20 (98.18-98.23)	97.63 (97.59-97.66)	98.46 (98.44-98.48)	0.973	0.977
70	0	20	10	80.5	96.12 (96.08-96.17)	99.62 (99.60-99.63)	99.51 (99.50-99.53)	98.48 (98.46-98.50)	0.978	0.980
70	0	30	0	72.5	97.06 (97.02-97.10)	98.28 (98.25-98.30)	97.84 (97.80-97.87)	98.45 (98.43-98.47)	0.974	0.977
70	10	0	20	75.5	96.27 (96.23-96.32)	98.05 (98.03-98.08)	97.42 (97.38-97.46)	98.40 (98.38-98.42)	0.968	0.973
70	10	10	10	80.5	95.80 (95.75-95.84)	99.74 (99.73-99.75)	99.67 (99.66-99.68)	98.48 (98.46-98.50)	0.977	0.979
70	10	20	0	78.5	96.31 (96.27-96.35)	99.56 (99.54-99.57)	99.45 (99.43-99.47)	98.48 (98.46-98.50)	0.979	0.981
70	20	0	10	79.5	95.52 (95.48-95.57)	98.44 (98.41-98.47)	98.06 (98.03-98.09)	98.30 (98.28-98.32)	0.968	0.971
70	20	10	0	79.5	95.77 (95.72-95.81)	99.12 (99.10-99.14)	98.93 (98.91-98.95)	98.37 (98.35-98.39)	0.973	0.976
70	30	0	0	73.5	92.42 (92.36-92.48)	97.12 (97.08-97.15)	96.40 (96.35-96.44)	97.34 (97.31-97.36)	0.944	0.950
80	0	0	20	83.5	93.72 (93.67-93.78)	99.05 (99.03-99.06)	97.90 (97.87-97.93)	98.05 (98.03-98.07)	0.958	0.973
80	0	10	10	84.5	96.01 (95.96-96.05)	99.15 (99.14-99.17)	98.22 (98.19-98.26)	98.48 (98.46-98.50)	0.971	0.981
80	0	20	0	81.5	96.13 (96.08-96.17)	99.09 (99.07-99.10)	98.12 (98.09-98.15)	98.44 (98.42-98.46)	0.971	0.981
80	10	0	10	85.5	95.17 (95.12-95.22)	99.04 (99.02-99.05)	98.01 (97.98-98.04)	98.38 (98.36-98.40)	0.966	0.978
80	10	10	0	85.5	95.67 (95.62-95.71)	99.61 (99.60-99.62)	99.21 (99.18-99.23)	98.45 (98.43-98.47)	0.974	0.983
80	20	0	0	80.5	92.09 (92.02-92.15)	97.90 (97.88-97.93)	95.73 (95.68-95.77)	97.76 (97.74-97.79)	0.939	0.959
90	0	0	10	91.5	92.92 (92.86-92.97)	99.50 (99.49-99.51)	98.30 (98.27-98.33)	98.17 (98.15-98.19)	0.955	0.979
90	0	10	0	90.5	95.33 (95.29-95.38)	99.49 (99.48-99.50)	98.33 (98.30-98.36)	98.44 (98.42-98.46)	0.968	0.985
90	10	0	0	90.5	90.99 (90.92-91.05)	98.91 (98.90-98.92)	96.40 (96.35-96.44)	98.02 (97.99-98.04)	0.936	0.970
100	0	0	0	94.5	95.25 (95.20-95.30)	89.22 (89.19-89.25)	65.68 (65.59-65.77)	93.60 (93.56-93.64)	0.778	0.903

Internal linkage APAC

The deterministic record linkage to create clusters of patient records processed 3,129,541 distinct CPF and CNS numbers, acquired from 24,174,912 records, from 2000 through 2012. After the time sequential deterministic linkage, we detected 1,834,932 UIDs.

The probabilistic record linkage performed during deduplication of UIDs further improved the clustering, linking an additional 92,783 (5,05%) of the original clusters with others by identifying like clusters but with one identified by a CPF and the other by a CNS, or identified by separate CNSs. Frequently three or more original clusters were linked together. This linkage brought the final number of UIDs to 1,742,149. We found 59,687 pairs classified as 100% confidence matches, 106,619 pairs classified as a match (score >= 65%). We classified as non-matches 12,810,135 pairs with score of over 50%.

Record linkage between APAC and AIH

The record linkage between the APAC Oncology database and cancer-related hospitalizations was performed to link the 1,742,149 individuals from APAC with the 2,330,677 patients in the SIH database. We found 821,640 record pairs classified as matches (score >= 72%) and 1,692,251 record pairs classified as non-matches based on confidence scores between 50% through 71%. At least one hospitalization record in the calendar range of 1 year before or after the start of chemo- or radiotherapy treatment was linked to 637,016 (36.56%) patients with APAC records. A total of 1,600,280 cancer-related hospitalizations were linked to APAC Oncology patients. On the other hand, 1,105,133 individuals in the APAC database, and 2,975,328 individuals in the AIH database presented no match. Information obtained from hospitalization was incorporated to adjust the date of first treatment for cancer, updating information with respect to the start of treatment for 401,327 individuals.

Record linkage between APAC and SIM

Our record linkage between the 1,742,149 individuals in the APAC Oncology database and cancer-related deaths in SIM (TABLE 3) produced 828,524 record pairs classified as matches (score >= 72%) and 8,260,681 record pairs classified as non-match having a score over 50%. We found at least one trustable match for 799,471 (45.89%) of APAC patients and no trustable match with a score over 50% for 1,413,639 individuals in APAC. Additionally, 1,167,939 records of the SIM database remained without any match in the APAC database.

TABLE 3 presents the time series of patient deaths recorded in the APAC database and totals including deaths recovered by probabilistic record linkage between APAC patients and SIM database records with cancer ICD-10 codes. We found 42,851 (89.2% of all notified deaths) of patients in the APAC Oncology database exclusively through this record linkage, resulting in an average underreporting of mortality in the APAC database of 88.7% (88.47% - 88.99%). Indeed, the percentage of deaths detected only through linkage rose 0.55% (0.52% - 0.58%) per year as we approached 2012, though this increase in underreporting between 2008 and 2012 was not statistically different from zero (p-value 0.90).

TABLE 3 – Impact of record linkage (RL) between APAC and SIM on the number of cancer-related deaths detected for patients in the APAC Oncology database (2001-2012)

Year	Deaths recorded in APAC N	Deaths after RL N	Information gain due to RL %	Deaths recorded in APAC and detected on RL N	APAC deaths detected by RL %
2001	-	27,096	-	-	-
2002	-	43,428	-	-	-
2003	-	47,782	-	-	-
2004	-	50,454	-	-	-
2005	-	57,249	-	-	-
2006	-	63,353	-	-	-
2007	-	66,299	-	-	-
2008	7,995	72,968	89.04	7,033	87.97
2009	9,099	79,304	88.53	8,060	88.58
2010	9,871	85,290	88.43	8,791	89.06
2011	10,213	88,573	88.47	9,152	89.61
2012	10,883	100,310	89.15	9,815	90.19
2013	-	17,365	-	-	-
Total	48,061	799,471	89.17	42,851	89.16

Discussion

In this study, we describe the process of building a cohort of patients treated with chemo- or radiotherapy for cancer financed by the SUS throughout Brazil over the period from 2000 to 2012. Our strategy of internal record linkage, initiating with deterministic and following with probabilistic deduplication allowed us to resolve the problem of non-unique patient identification keys over the years and to create clusters of records represented by a unique identifier which permit epidemiological analyses of the cohort of patients. We then linked in information from hospitalization and death records to improve the quality and scope of information present.

Since 1995, centers for data linkage have increased in several of the world´s nations, including Australia, the UK, Scotland, Germany and Canada [24–34]. Other regions, without the availability of similar data linkage facilities, also perform high quality studies using record linkage. A few independent studies showed overall sensitivities in the linkage of administrative health system databases of between 74% to 98%, specificities ranging from 99% to 100%, and precision from 93% to 99% [27]. Australian research centers have excellent record linkage quality statistics on population-based health databases, revealed F-measures between 0.95 and 0.99, accuracy from 93% to 99.9%, and recall from 91% to 99.7% [26].

Quality and accuracy of the probabilistic record linkage process are related to several factors such as the 1) coverage of the information system; 2) the quality of the completeness of variables of identification – selected according to their discriminatory capacity to compare the pairs using measures of similarity [35]; and 3) the appropriateness of the model used to determine the confidence scores which permit matching of pairs of records [35,36].

The degree of completeness of the variables we used in the record linkage was high. may have a considerable impact on the probability of detecting a true record pair. Following suggestions from the literature [36,37], we overcame part of residual incompleteness (e.g.,

invalid, ignored, and missing values) by assigning a null value to linkage variables that presented missing values. Additionally, the quality of the information present in chemo-radiotherapy APAC records improved considerably from the beginning of the series to the latter years between 2008 and 2012.

The SIH data we used for the first years of our series has been classified as having low reliability, fundamentally with respect to the diagnostic variables coded in ICD-10[38,39]. The government agency RIPSA estimated that, in 2000, the ratio of recorded deaths and estimated was 91%, 93.2% in 2005, and 94.2% in 2010 nationwide. SIM data, although presenting quantitative and qualitative limitations that indicate the need for further actions, have improved greatly in recent years. The Ministry of Health, during the period evaluated by the study, made several efforts to improve and guarantee the quality of the SIM - for example, intensifying the training of the health professionals involved, particularly the coders of the basic cause of death, as well as correction for causes garbage codes of ICD-10, and alerting municipalities of the need to guarantee the trustworthiness of this database[40–45].

Queiroz et al. [46] evaluated reproducibility and validity of death certificates from SIM of women at reproductive age in Belém do Pará city, Brazil. They estimated an agreement of classification of 93% (kappa = 0.92) between ICD-10 chapters, and a positive predictive value ranging from 83.7% to 100%, in the diverse chapters. Classification disagreements in the neoplasms chapter occurred less than in the other disease chapters – originally neoplasms were 29.9% of the cases in the SIM classification, and rose to 30.5% after reclassification.

Although through record linkage between APAC Oncology and SIM we substantially improved the quality of mortality information, we note that only approximately 90% of deaths already reported in APAC were detected as true pairs in this linkage. However, the undetected cases (losses) may be due not to problems of record linkage, but rather to our original filtering of SIM registers to exclude records not having malignant neoplasms ICD-10. We are thus unable

to state the fraction of undetected deaths related to this classification mismatch as opposed to the problems shown by authors[46].

Freire et al. [47], found that 41.2% of the individuals present in the chemo-radiotherapy APAC could be linked to a death record in SIM over the period from 2000 through 2005 in the state of São Paulo. Despite all possible methodological limitations, 45.9% of the individuals present in the APAC Oncology linked to a SIM death record over the period 2001 through 2012, meaning a 4.7% improvement in this aggregation of mortality information.

The final model selected for record linkage, when applied to the APAC Oncology internal linkage for patient identification (deduplication), presented a satisfactory quality, with an area under the ROC curve of 99.56%, sensitivity of 97.88%, predictive positive value of 99.50%, F-measure of 0.987, and accuracy of 0.983. These results evaluated only pairs with scores greater than or equal to 50% of confidence. We evaluated these measures taking into account all the scores of pairs (0% -100%), leading to a specificity to 99.99% for all sets of weights tested (high number of true non-matches record pairs). Therefore, we limited the pair storage to files of $\geq$ 50% confidence scores due to computational and disk storage cost – the final file size with record linkage results was higher than 25Gb per APAC year by SIM year comparison. The distribution curve of true pairs, identifiable by previous known CNS and CPF numbers, had zero density in the confidence score region 0% up to 49%.

A heterogeneity of methods among record linkage studies exposes a lack of standardization with respect to the communication of findings, of missing information, and in the most cases, of the quality of linkage to allow more accurate comparison of findings. Thus, we only compared studies with record linkage measures indicating quality. Among the studies using APAC data, the most similar study to ours, Freire et al. [47] carried out a record linkage of the databases of chemo-radiotherapy APAC, SIM and SIH over the period from January 2000 thought December 2004, for records of the state of Rio de Janeiro in Brazil. They found, for the

internal linkage of chemo-radiotherapy APAC, a sensitivity of 99.44% (99.79% - 99.98%) and a specificity of 100% (99.9% - 100%).

Other studies related to cancer information systems, conducted in Brazil, found the following results: Freire et al. [48] applied record linkage on Brazilian Cervical Cancer Information System (SISCOLO) in order to identify all women in the system, and found a sensitivities above 90%, and specificities and precisions near 100%. Peres et al [49] explored a cutoff point for identifying the true pairs on probabilistic record linkage of Population-based Cancer Registry of São Paulo (registered from 1997 through 2005) with chemo-radiotherapy APAC (August 2003 through December 2007) database, restricted to the municipality of São Paulo. The researchers stratified the analysis by cancers topography and detected an area under the ROC curve ranging from 94.7% to 99%, with sensitivity ranging from 73.7% to 96.7% and specificity ranging from 98.5% to 99.4%.

The EM algorithm is the method generally applied to determinate the score used for classification of pairs of records in true matches, or false matches probabilities, on similar studies mentioned above. The findings concerning weights of record linkage variables demonstrated substantial difference among empirical weights approach and weights of EM algorithm. We believe that the difference may be due to the tendency of the EM algorithm to yield inaccurate estimates when a significant number of field values are missing[36].

Some limitations of the current study should be mentioned. Record linkage variable weights were estimated from the APAC Oncology data and then readjusted for linkages between APAC Oncology with SIH, and APAC Oncology with SIM. We know that the diversity of information completeness among the databases (Table 1), and the fact that APAC, SIH and SIM do not share all the record linkage variables used during the process of weight estimation, may imply a loss of quality on record linkage. Although this may seem to be a weakness of the linkage strategy, the EM algorithm when computed over the linkages between APAC Oncology with SIH,

and between APAC Oncology with SIM, showed the same discrepant behavior for the weights of linkage variables to create the records pair confidence score. This leads us to believe that this approach retains its properties of quality. Another limitation is the fact that not all cancers in Brazil receive treatment financed by the SUS. The fraction is hard to estimate, as databases describing these treatments have not been available for analysis. Private health plans are now estimated to cover 27.9% of the population [50], although many of these do not provide for more sophisticated cancer treatments. Future work should incorporate these data in analyses of the overall treatment of cancer in Brazil.

We elected Levenstein's distance to access the similarity of strings of linkage variables based on their simplicity, wide use in other works, and availability in FRIL software. However, for use in an automatic matching scheme, Cohen et al. [20] suggest that other measures of similarity, such as Jaro-Winkler's distance, would offer a better result in linking pairs of names of the same entity.

The blocking and comparison method applied in this study were not intended to optimize the record linkage execution time. We prime for a better true pairs detection strategy instead, considering limitations of information completeness, and characteristics related to municipality of residence. This choice increases computational cost, both in resources and in processing time, although the use of new effective blocking methods or other studies can be easily implemented in the record linkage process.

In the context of the epidemiological analysis of the data, the chemo-radiotherapy APAC database files, over the period between 2000 through 2007, provided by DATASUS did not contain information of patient’s mortality and reason for leaving the treatment of cancer. According to DATASUS documents[51], this information is part of the database for this period. If the study had this information in addition to epidemiological exploration, this could be used to monitor the quality of mortality linkage.

This study, to our knowledge, is the first effort, using record linkage, and a long period of time as 2000 thought 2012, to create a cohort of cancer patients from large databases of the Brazilian national health system.

Throughout this study, we tested various approaches to all sections of the record linkage process. Some of these approaches, while fairly simple, have proven to be quite useful and productive for improving the quality and accuracy of the linkage strategy as a whole. Some not so widespread in the literature on record linkage and deserve to be highlighted, as awareness of them, whose implementation may come at little or no additional cost and may have a great impact on results, could alter decisions as to the choice of the linkage strategy made by a researcher who faces similar decisions:

- The data have their own characteristics and knowing their particularities, even with respect to differences over time is fundamental to define an adequate linkage strategy.
- The completeness of the record linkage variables, and the entropy of their information, considerably affect the weights associated with the information used in linkage and the cut-off point for match and non-match pairs. Depending on the data, over time these optimal weights and cut-off points may be different. Thus, exploiting the data and time dimension of them is critical.
- Record linkage is a trade-off between time spent and accuracy, and as such requires obtaining a balance for each specific situation.

Record linkage allows us to create a database capable of monitoring the flow of cancer patients whose treatment was financed by the SUS. The APAC Oncology database provides information such as the number and distribution of cases by type of cancer, cancer stage, an outline of patients (age, sex, race/color), municipality of residence and treatment (migration), time elapsed between diagnosis and treatment, survival analysis, type of treatment performed.

The APAC information system covers other diseases treated by SUS. Thus, the methodology can derive tools for monitoring procedures for other diseases in Brazil, based on other APAC database series than cancer. In addition, this methodology may be applied in countries where similar, difficult to analyze duo non-common identifiers, administrative databases have been constructed but never adequately analyzed.

Conclusion

APAC Oncology database, the result of a few record linkage, concentrating important epidemiological information can be useful to generate evidence to inform policy and services. Naturally, there are some limitations: record linkage of existing administrative datasets are a powerful and efficient way of generating new information, but in function of errors related to data processing before, during, and after linkage, bias can be generated. Regardless those issues, SUS-APAC oncology database may be an important tool for the Ministry of Health, to construct indicators and monitoring, in accordance with Brazilian Strategic Action Plan to Combat Chronic Non-communicable Diseases and the global targets set to confront these diseases by 2025.

Declarations

Acknowledgments

We thank DATASUS-Rio for making available data used in this study, especially Guido Rafael Le Senechal Salatino, Othon Murilo Tupinambá Lourenço, Norberto Peçanha da Silva and Haroldo Lopes dos Santos; Antony Stevens for sharing their experience with Record Linkage, programming, data encryption, and data security arrangements.

Funding/Sponsorship

This work was conducted during a scholarship supported by the International Cooperation Program CAPES/STICAMSUD at the Emory University. Financed by CAPES – Brazilian

Federal Agency for Support and Evaluation of Graduate Education within the Ministry of Education of Brazil.

The funders had no role in study design, data collection, analysis, decision to publish, or preparation of the manuscript.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

IP, and BBD developed the study content and design. IP performed the record linkage, electronic matching, created the base analytic dataset, analyzed and interpreted the data, created the tables and figures, and wrote the first draft of the Introduction, Methods, and Results. IP and BBD all edited and approved the final version. BBD oversaw the research process. XX, YY, and ZZ consistently supplied critical inputs during review of the drafts. All authors contributed to interpreting the data, critically reviewed the drafts, and approved the final manuscript.

References

1. Fitzmaurice C, Allen C, Barber RM, Barregard L, Bhutta ZA, Brenner H, et al. Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life-years for 32 Cancer Groups, 1990 to 2015: A Systematic Analysis for the Global Burden of Disease Study. JAMA Oncol. 2017;3:524–48.

2. Instituto Nacional de Câncer. Estimativa 2016: incidência de câncer no Brasil [Internet]. Rio de Janeiro: INCA; 2015. Available from: www.inca.gov.br/estimativa/2016/estimativa-2016-v11.pdf

3. Schmidt MI, Duncan BB, e Silva GA, Menezes AM, Monteiro CA, Barreto SM, et al. Chronic non-communicable diseases in Brazil: burden and current challenges. The Lancet. 2011;377:1949–61.

4. Instituto Brasileiro de Geografia e Estatística, editor. Pesquisa nacional de saúde, 2013: acesso e utilização dos serviços de saúde, acidentes e violências: Brasil, grandes regiões e

unidades da Federação. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística-IBGE; 2015.

5. Farmer P, Frenk J, Knaul FM, Shulman LN, Alleyne G, Armstrong L, et al. Expansion of cancer care and control in countries of low and middle income: a call to action. Lancet Lond. Engl. 2010;376:1186–93.

6. Alwan A, MacLean DR, Riley LM, d'Espaignet ET, Mathers CD, Stevens GA, et al. Monitoring and surveillance of chronic non-communicable diseases: progress and capacity in high-burden countries. The Lancet. 2010;376:1861–8.

7. Sørensen HT, Lash TL. Use of administrative hospital registry data and a civil registry to measure survival and other outcomes after cancer. Clin. Epidemiol. 2011;3:1.

8. Prestes IV, de Moura L, Duncan BB, Schmidt MI. A national cohort of patients receiving publicly financed renal replacement therapy within the Brazilian Unified Health System. The Lancet. 2013;381:S119.

9. Moura L de, Prestes IV, Duncan BB, Schmidt MI. Building a national database of patients receiving dialysis on the Brazilian Unified Health System, 2000-2012. Epidemiol. E Serviços Saúde. 2014;23:227–38.

10. Winkler WE. Matching and record linkage. Wiley Interdiscip. Rev. Comput. Stat. 2014;6:313–25.

11. Vick R, Huynh L. The Effects of Standardizing Names for Record Linkage: Evidence from the United States and Norway. Hist. Methods J. Quant. Interdiscip. Hist. 2011;44:15–24.

12. Herzog TN, Scheuren FJ, Winkler WE. Data Quality and Record Linkage Techniques. Springer Science & Business Media; 2007.

13. Blakely T, Salmond C. Probabilistic record linkage and a method to calculate the positive predictive value. Int. J. Epidemiol. 2002;31:1246–52.

14. Dusetzina SB, Tyree S, Meyer A-M, Meyer A, Green L, Carpenter WR. An Overview of Record Linkage Methods [Internet]. Agency for Healthcare Research and Quality (US); 2014 [cited 2017 Apr 20]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK253312/

15. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143:29–36.

16. Jurczyk P, Lu JJ, Xiong L, Cragan JD, Correa A. FRIL: A Tool for Comparative Record Linkage. AMIA. Annu. Symp. Proc. 2008;2008:440–4.

17. Murray JS. Probabilistic Record Linkage and Deduplication after Indexing, Blocking, and Filtering. ArXiv160307816 Stat [Internet]. 2016 [cited 2017 Apr 20]; Available from: http://arxiv.org/abs/1603.07816

18. Mamun A-A, Mi T, Aseltine R, Rajasekaran S. Efficient sequential and parallel algorithms for record linkage. J. Am. Med. Inform. Assoc. JAMIA. 2014;21:252–62.

19. Levenshtein VI. Binary codes capable of correcting deletions, insertions, and reversals. Sov. Phys. Dokl. [Internet]. 1966 [cited 2015 Feb 13]. p. 707–710. Available from:

https://gitlab.doc.ic.ac.uk/wm613/inidividual-project/raw/4fbe43f863b229b50ca13bf5f59eb029ad71f6b6/reading/litreview/levenshtein66.pdf

20. Cohen W, Ravikumar P, Fienberg S. A comparison of string metrics for matching names and records. KDD Workshop Data Clean. Object Consol. [Internet]. 2003 [cited 2015 Jan 25]. p. 73–78. Available from: https://www.cs.cmu.edu/afs/cs/Web/People/wcohen/postscript/kdd-2003-match-ws.pdf

21. TrueCrypt [Internet]. [cited 2017 Apr 16]. Available from: http://truecrypt.sourceforge.net/

22. Miao Q x. Research and analysis on Encryption Principle of TrueCrypt software system. 2nd Int. Conf. Inf. Sci. Eng. 2010. p. 1409–12.

23. VeraCrypt download | SourceForge.net [Internet]. [cited 2017 Apr 16]. Available from: https://sourceforge.net/projects/veracrypt/

24. Zealand SN. Integrated Data Infrastructure and prototype. Stat. N. Z. Wellingt. 2012;

25. Zealand SN. Linking methodology used by Statistics New Zealand in the Integrated Data Infrastructure project. Technical report. Available from www. stats. govt. nz.[Cited on page 4.]; 2014.

26. Boyd JH, Randall SM, Ferrante AM, Bauer JK, McInneny K, Brown AP, et al. Accuracy and completeness of patient pathways – the benefits of national data linkage in Australia. BMC Health Serv. Res. 2015;15:312.

27. Silveira DP da, Artmann E. Accuracy of probabilistic record linkage applied to health databases: systematic review. Rev. Saude Publica. 2009;43:875–82.

28. Fleming M, Kirby B, Penny KI. Record linkage in Scotland and its applications to health research. J. Clin. Nurs. 2012;21:2711–21.

29. Holman CDJ, Bass JA, Rosman DL, Smith MB, Semmens JB, Glasson EJ, et al. A decade of data linkage in Western Australia: strategic design, applications and benefits of the WA data linkage system. Aust. Health Rev. 2008;32:766–77.

30. Lawrence G, Dinh I, Taylor L. The Centre for Health Record Linkage: A New Resource for Health Services Research and Evaluation. Health Inf. Manag. J. 2008;37:60–2.

31. Kendrick S, Clarke J. The Scottish record linkage system. Health Bull. (Edinb.). 1993;51:72.

32. Gill L. OX-LINK: the Oxford medical record linkage system. Record Linkage Techniques, http://www. fcsm. gov/working-papers/gill. pdf; 1997.

33. International Journal of Population Data Science. Data linkage centres | www.ipdln.org [Internet]. 2017 [cited 2017 Aug 11]. Available from: http://www.ipdln.org/data-linkage-centres

34. Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staa T, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). Int. J. Epidemiol. 2015;44:827–36.

35. Jaro MA. Probabilistic linkage of large public health data files. Stat. Med. 1995;14:491–498.

36. Bauman GJ. Computation of weights for probabilistic record linkage using the EM algorithm. 2006;

37. Karmel R, Anderson P, Gibson D, Peut A, Duckett S, Wells Y. Empirical aspects of record linkage across multiple data sets using statistical linkage keys: the experience of the PIAC cohort study. BMC Health Serv. Res. 2010;10:41.

38. Lima CR de A, Schramm JM de A, Coeli CM, Silva MEM da. Review of data quality dimensions and applied methods in the evaluation of health information systems. Cad. Saúde Pública. 2009;25:2095–109.

39. Brasil, RIPSA. Índice A.18 - Razão entre óbitos informados e estimados [Internet]. 2015 [cited 2017 Aug 12]. Available from: http://tabnet.datasus.gov.br/cgi/idb2011/a18.htm

40. Brasil. Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Análise de Situação de Saúde*. Portaria nº 653/GM de 28 de maio de 2003 [Internet]. 2003. Available from: https://www.google.com.br/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0ahUKEwie0ta2scDVAhWBgZAKHXOxAo0QFggrMAA&url=http%3A%2F%2Fwww.cvs.saude.sp.gov.br%2Fzip%2FPortaria%2520GM%2520MS%2520n%25C2%25BA%2520653%2C%2520de%252028maio03.pdf&usg=AFQjCNF0oWvSkz5OP76qoyyeQdmVF2KVYQ

41. Gomes F de BC. Sistema de informações sobre mortalidade: considerações sobre a qualidade dos dados. Inf. Epidemiológico Sus. 2002;11:5–6.

42. Brasil. Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Análise de Situação de Saúde*. PORTARIA Nº 1.119, DE 5 DE JUNHO DE 2008 [Internet]. 2008 [cited 2017 Aug 5]. Available from: http://bvsms.saude.gov.br/bvs/saudelegis/gm/2008/prt1119_05_06_2008.html

43. Rattner D, Rabello Neto D, Lansky S, Vilela M, Bastos M. Mortalidade do adulto no Brasil: taxas de mortalidade segundo sexo, causas e regiões, 2010. Saúde Bras. 2011 Uma Análise Situaç. Saúde E Vigilância Saúde Mulher. 2012.

44. França E, Teixeira R, Ishitani L, Duncan BB, Cortez-Escalante JJ, Neto M, et al. Ill-defined causes of death in Brazil: a redistribution method based on the investigation of such causes. Rev. Saúde Pública. 2014;48:671–81.

45. Stevens A, Schmidt MI, Duncan BB. Information–processing methods for mortality surveillance in the presence of varying levels of completeness and ill–defined codes of causes of death – the case of Brazil. Popul. Health Metr. 2014;12:24.

46. Queiroz RC de S, Mattos IE, Monteiro GTR, Koifman S. Reliability and accuracy of oral cancer as the reported underlying cause of death in the Municipality of Rio de Janeiro. Cad. Saúde Pública. 2003;19:1645–53.

47. Freire SM, Souza RC de, Almeida RT de. Integrating Brazilian health information systems in order to support the building of data warehouses. Res. Biomed. Eng. 2015;31:196–207.

48. Freire SM, Almeida RT de, Cabral MDB, Bastos E de A, Souza RC, Silva MGP da. A record linkage process of a cervical cancer screening database. Comput. Methods Programs Biomed. 2012;108:90–101.

49. Peres SV, Latorre M do RD de O, Michels FAS, Tanaka LF, Coeli CM, Almeida MF de. Determinação de um ponto de corte para a identificação de pares verdadeiros pelo método probabilístico de linkage de base de dados. Cad. Saúde Coletiva. 2014;22:428–36.

50. Malta DC, Stopa SR, Pereira CA, Szwarcwald CL, Oliveira M, Reis AC dos, et al. Private Health Care Coverage in the Brazilian population, according to the 2013 Brazilian National Health Survey. Ciênc. Amp Saúde Coletiva. 2017;22:179–90.

51. Arquivo(s) de APAC para download. [Internet]. [cited 2015 Jan 27]. Available from: http://sia.datasus.gov.br/documentos/listar_ftp_apac.php

ARTIGO 2

Tratamento de cânceres potencialmente curáveis no Sistema Único de Saúde, 2000 a 2012: incidência e tempo decorrido entre diagnóstico e início do tratamento

Treatment of potentially curable cancers financed through the Brazilian National Health System, 2000 to 2012: incidence and time from dignosis to initial treatment

Isaías Valente Prestes, Doutorando em Epidemiologia pela UFRGS.

Bruce Bartholow Duncan, Federal University of Rio Grande do Sul, Postgraduate Program in Epidemiology

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL (UFRGS)

A ser enviado ao BMC Cancer - https://bmccancer.biomedcentral.com/

TITLE

Treatment of cancers financed through the Brazilian National Health System, 2000 to 2012 – Incidence and time elapsed from diagnosis to initiation of treatment

ABSTRACT

BACKGROUND - Cancer is one of the leading causes of morbidity and mortality worldwide. Despite treatment of over one million cancers financed by the Brazilian national health system (Sistema Único de Saúde, or SUS) over recent years, little has been reported characterizing this treatment.

OBJECTIVE – To describe frequency, trends, and punctuality of publicly-financed chemo- and radiotherapy of cancer within the SUS, with emphasis on a subset of potentially treatable and/or frequent cancers.

RESULTS: Over the period 2001-2012, we observed an age-standardized increase in these therapies which was significantly greater than that of estimated increases in cancer incidence. Among women, treatment of breast (27.5%), bronchus and lung (45.8%), and colorectal cancer (61.2%) expanded while treatment of thyroid (-16.7%) and cervical cancers (-21.2%) decreased. Among men, we detected significant increases in initial chemo- or radiotherapy for prostate (108.5%), colorectal (72.8%), breast (23.2%) and testicular (36.5%) cancer. Among patients receiving treatment for the specific cancers studied from 2008 to 2012, 62.7% (IC95% 62.5; 62.9%) received initial treatment within 60 days after diagnosis, this number being unchanged over the period.

CONCLUSION: Publicly-financed cancer treatment has expanded considerably since 2001 in Brazil, but the waiting time between diagnosis and treatment is long and the fraction of patients treated within 60 days is basically unchanged over recent years.

Key words: Cancer; Health Information Systems; Epidemiological Surveillance; Record Linkage.

INTRODUCTION

Cancer is one of the leading causes of morbidity and mortality worldwide, with an estimated 17.5 million incident cancer cases, and 8.7 million cancer deaths globally in 2015[1]. According to the Brazilian National *Cancer* Institute (INCA), approximately 490,000 new cases of cancer were diagnosed in 2011 and 596,000 new cases in 2016, implying an increase of 22% over 5 years. The ranking of the most frequent is similar to that of high income countries: lung, breast, colorectal, stomach, prostate, liver, cervical, and esophageal, and has remained unchanged recently[2].

A global call to action to confront the non-communicable diseases (NCDs), including cancer, in low-income and middle-income countries (LMICs) was initiated by the WHO in 2005[3], and calls to action by international agencies, academic institutions, and nongovernmental organizations and associations accelerated over the following decade [4–7]. With respect to cancer, focus has been placed internationally not only on prevention through risk factor control, but also on improving outcomes for those with treatable and potentially curable cancers whose natural history permits detection in time for effective actions. Such interventions should be accessible to patients and integrated into national health systems. Expansion of cancer care in LMICs should include increasing access to drugs for treatment and palliation; expansion of coverage for preventive and diagnostic services, including vaccines; and development and implementation of innovative health-care delivery options to support expansion of cancer care[8].

Estimates of the extent and nature of interventions against cancer in individual countries and regions are necessary to inform national cancer control strategies. Cancer registries and administrative health care databases can provide information at a population level given their representativeness and large numbers of records of medical care performed in hospitals and healthcare systems [9,10].

In order to promote access to cancer information in the Brazilian National Health System (SUS), we created a database for epidemiological analyses of oncology treatments financed by the SUS[9], which, according to information from the 2013 National Health Survey, is the exclusive source of care for 72.1% of Brazilians[11]. This database contains patient histories of all cancers receiving publicly-financed chemo- or radiotherapy for cancer from 2000 to 2012. A product of record linkage, it aggregates information on chemo- and radiotherapy of the high complexity procedures authorization system (Autorização de Procedimentos Ambulatoriais de Alta Complexidade/Custo, or APAC), as well as information present about these patients in the national hospital (Sistema de Informações Hospitalares or SIH) and mortality (Sistema de Informação sobre Mortalidade, or SIM) information systems.

The objective of this report is to describe, using this oncology database, the epidemiology of cancers deemed to be potentially treatable and/or frequent whose care was financed by the SUS.

MATERIAL AND METHODS

We carried out a descriptive study of the publicly-financed treatment of these cancers over the period of 2000 through 2012. The oncology database was created by linkage of the data of patients in the APAC database, which records chemotherapy and radiotherapy for reimbursement purposes [9]. Thus, all patients who received one or both treatments were included. We also performed probabilistic record linkage of APAC records with hospitalization information from the SUS Hospital Information System (SIH/SUS), over the period of 2004 through 2012. This allowed us to include publicly-financed initial care and surgery treatments and, in some cases, obtain a more accurate date of initial treatment.

We selected cases on the basis of codes of the International Classification of Diseases, ICD-10 (C00 to C97 and D37 to D48) listed in the APAC variables of primary, secondary, or associated cause of treatment or as the assigned cause for oncological procedures, except when

they related to benign conditions (0304010014 and 0304010235). Using all this information, we classified the treated cancers based on the ICD-10 code of the malignant neoplasm listed as the primary cause as breast (C50), bronchus or lung (C34), cervical (C53), colorectal (C18, C19, C20), Hodgkin's lymphoma (C81), non-Hodgkin's lymphoma (C82, C83, C85), lymphoblastic leukemia (C91), myeloid leukemia (C92), prostate (C61), testicular (C62), thyroid (C7 3), or as any other malignant neoplasm (other ICD-10 Chapter II, Letter C codes). The specific cancers listed above were chosen, after consultation with the Ministry of Health, based on a combination of their frequency and their potential for possible cure with treatment [8]. APAC instructions oriented the indication of cancer stage following the TNM Classification of Malignant Tumors of the International Union (UICC)[12,13], with tumors being staged from 0 to IV.

We expressed crude and adjusted initial and ongoing treatment, per 100,000 inhabitants/year, using population data from the 2000 and 2010 demographic censuses and official interpolations for intercensus years [14]. As we could not distinguish new cases from prevalent cases in 2000, we defined new treatments as those not appearing previously within APAC from 2001 onwards. Age-adjusted rate calculations are based on the WHO Standard Population, which is stratified into 19 age groups[15].

As of 2012, Brazilian law 12732 requires that cancer patients must receive initial treatment within 60 days of diagnosis. To characterize adherence to this mandate, we evaluated time elapsed, in days, between diagnosis and start of treatment based on information from APAC and hospitalization databases. We chose the period from 2008 through 2012 for these analyses as prior to 2008 date of diagnosis was missing for a large fraction of cases. Over the 2008-2012 period, 953,465 patients had APAC records indicating treatment whose primary cause was an ICD-10 coded malignant neoplasm. We included, based on APAC information, only the 230,431 cases with a valid date of diagnosis and start of chemo- or radiotherapy, and whose

record was the first request for reimbursement for cancer treatment in the SUS over the 2000-2012 period.

We performed joinpoint regression analysis (National Cancer Institute, Bethesda, MD), version 4.5.0.0 of May 2017 [16,17] over the period 2001 to 2012 to identify incidence trends and possible moments where a significant change in a trend occurred, permitting a maximum of 3 change points. We characterized annual percentage change (APC) for each time segment using the natural logarithm of the ratio of the treatment rates of different years [18]. We defined the final model based on significance tests obtained using the Monte Carlo permutation method. In describing trends, the terms "significant increase" or "significant decrease" mean that the slope of the trend was significant different from zero (p-value < 0.05). For nonsignificant trends, we apply the term "stable". Moreover, we performed Kruskal–Wallis test by ranks to evaluate differences in distribution among groups, and Mood's test of median to detect differences of median among groups. We performed data analysis with the Statistical Analysis System (SAS® 9.4), and produced tables and plots design in Excel® 2013 for Windows.

This study was approved by the Ethics in Research Committee of the Hospital de Clínicas de Porto Alegre on 10/03/2010 (No. 100056). As analyses are based on surveillance databases from the Ministry of Health, no patient consent was necessary. All those with access to data signed agreements to maintain the confidentiality of the information.

RESULTS

From 2000-2012, 1,742,149 individuals were identified as receiving cancer chemo- or radiotherapy financed by the SUS. Among these individuals, 1,570,311 initiated chemo- or radiotherapy for any malignant cancer (C00-C99) over the period of 2001 through 2012 (Table 1). Women represent 53.8% of these individuals. In terms of the type of treatment, 42.7% received only chemotherapy, 24.1% only radiotherapy, and 33.2% chemo- and radiotherapy. Additionally, 9.5% had at least one surgical procedure financed by the SUS during the first year

of treatment. More than 50% of initial treatments were for advanced stage cancer (stages III or IV), with 17.1% of women and 29.4% of men presenting with stage IV disease.

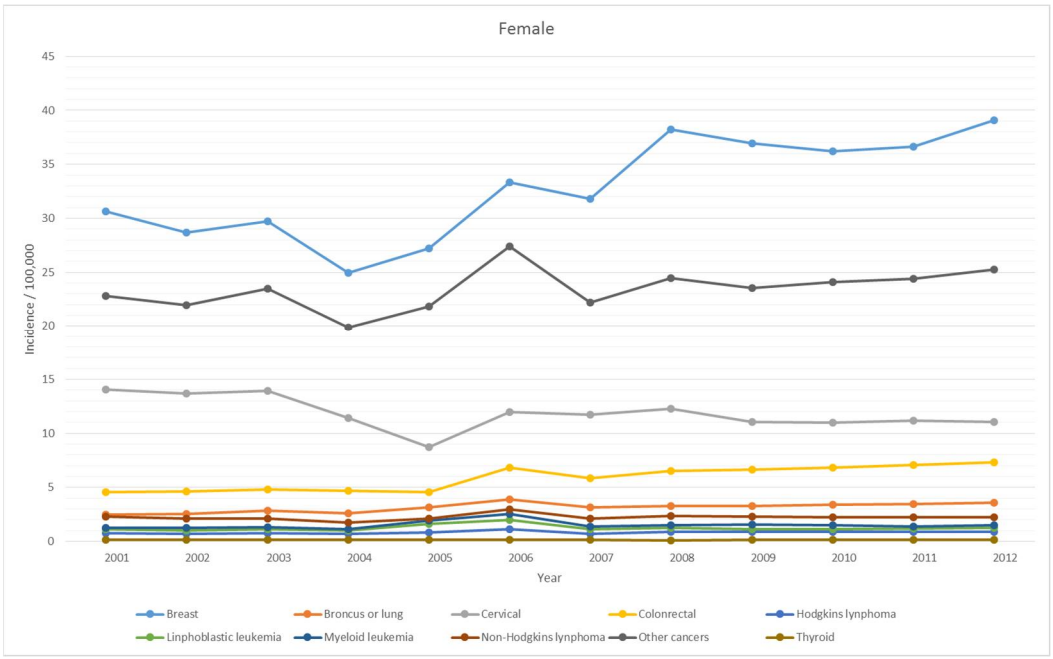
Incidence of treatment by type of cancer over 2001 though 2012 is presented in Figure 1 and Table 3. Over this period, the number initiating treatment increased 63.8%, from 104,019 in 2001 to 170,435 in 2012. Among women, breast and cervical cancers were by far the most common cancers receiving treatment (Figure 1A); among men, prostate cancer was the most common (Figure 1B). In comparison, the Brazilian population increased 13.3% over this same period. Initial chemotherapy treatments increased 87.2% (from 60,400 to 113,042 cases), while all cases receiving chemotherapy (initial and ongoing) expanded 140.2% (from 136,360 to 327,482). Over the same period, initial radiotherapy increased 31.7% (from 43,582 to 57,393) and all radiotherapy 39.9% (from 53,154 to 74,835).

Similar data for both initial and ongoing treatment of specific cancers are shown in Supplementary Figure 1. Breast cancer for females and prostate cancer for males are by far the most frequent cancers receiving treatment. Rates for both demonstrated major increases over the period. For males, treatment for other cancers – principally esophagus (10.3% of total), stomach (9.7%), larynx (9.2%), bladder (4.8%), brain (4.4%), and other and unspecified malignant neoplasm of skin (6%) – increased only slightly in relative terms. However, given the magnitude of therapy of these cancers, this increase was not small in absolute terms. Other cancers receiving initial or ongoing treatment for men and women, taken together, were principally stomach (8.1%), esophagus (7.7%), larynx (6.2%), ovary (5%), uterus (4.8%), and other and unspecified malignant neoplasm of skin (6.6%).

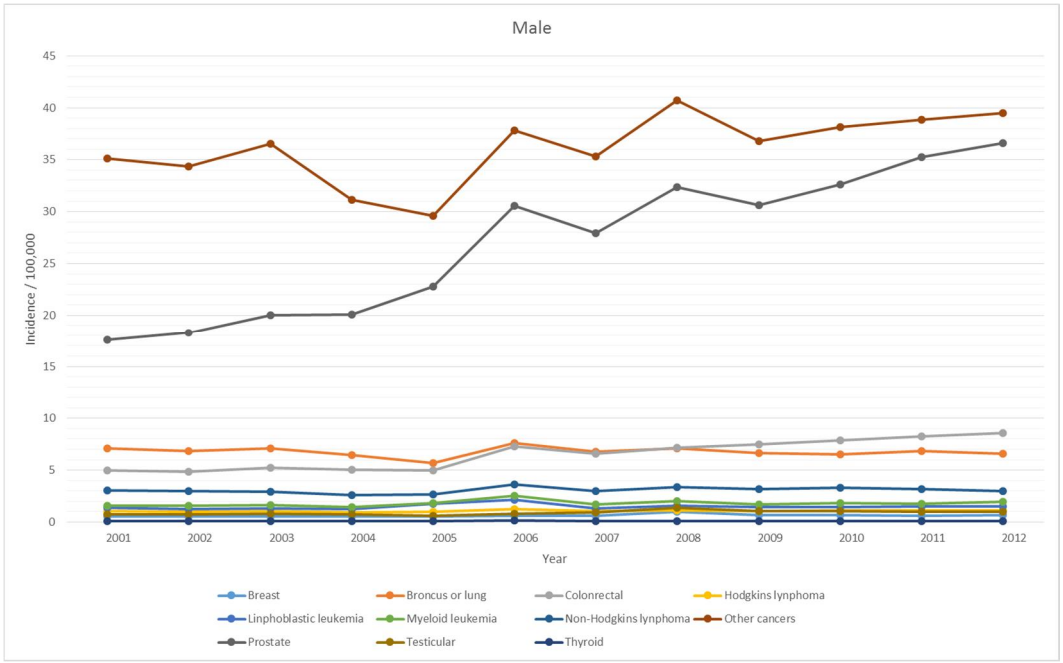
TABLE 1 – Sociodemographic and clinical characteristics of patients receiving their first publicly-financed cancer treatment in Brazil, 2001-2012, N=1,570,311.

Characteristic	Total		Female		Male	
	n	%	n	%	n	%
Region						
Central-west	95,155	6.1	51,820	6.1	43,335	6.0
North	60,751	3.9	36,819	4.4	23,932	3.3
Northeast	324,531	20.7	187,410	22.2	137,121	18.9
South	325,558	20.7	168,346	19.9	157,212	21.7
Southeast	764,316	48.7	401,157	47.4	363,159	50.1
Age group						
0 to 4	18,945	1.2	8,657	1.0	10,288	1.4
5 to 9	13,760	0.9	5,830	0.7	7,930	1.1
10 to 14	13,561	0.9	6,043	0.7	7,518	1.0
15 to 19	15,998	1.0	6,819	0.8	9,179	1.3
20 to 24	18,889	1.2	9,135	1.1	9,754	1.3
25 to 29	26,679	1.7	15,653	1.9	11,026	1.5
30 to 34	40,217	2.6	27,591	3.3	12,626	1.7
35 to 39	62,775	4.0	45,930	5.4	16,845	2.3
40 to 44	99,556	6.3	72,264	8.5	27,292	3.8
45 to 49	139,683	8.9	95,887	11.3	43,796	6.0
50 to 54	166,099	10.6	103,310	12.2	62,789	8.7
55 to 59	181,387	11.6	102,498	12.1	78,889	10.9
60 to 64	189,277	12.1	96,244	11.4	93,033	12.8
65 to 69	187,094	11.9	84,774	10.0	102,320	14.1
70 to 74	168,700	10.7	69,721	8.2	98,979	13.7
75 to 79	125,834	8.0	49,887	5.9	75,947	10.5
80 to 84	65,438	4.2	27,711	3.3	37,727	5.2
85 to 89	26,506	1.7	12,322	1.5	14,184	2.0
90+	9,285	0.6	4,918	0.6	4,367	0.6
Missing	628	0.0	358	0.0	270	0.0
Race/Color*						
White	352,961	63.7	184,512	63.3	168,449	64.2
Black	30,410	5.5	14,658	5.0	15,752	6.0
Brown	160,281	28.9	86,837	29.8	73,444	28.0
Yellow (Asian)	9,788	1.8	5,346	1.8	4,442	1.7
Native Brazilian	379	0.1	196	0.1	183	0.1
Invalid information	1,661	0.3	717	0.2	944	0.4
Lacking information	215,331	38.9	115,951	39.8	99,380	37.9
Cancer type						
Breast	380,284	24.3	359,736	42.8	20,548	2.8
Bronchus or lung	82,615	5.3	29,335	3.5	53,280	7.4
Cervical	118,036	7.5	118,036	14.1		
Colorectal	112,998	7.2	57,568	6.9	55,430	7.7
Hodgkin's lymphoma	19,559	1.3	8,719	1.0	10,840	1.5
Lymphoblastic leukemia	28,564	1.8	12,182	1.5	16,382	2.3
Myeloid leukemia	31,592	2.0	14,937	1.8	16,655	2.3
Non-Hodgkin's lymphoma	47,064	3.0	20,844	2.5	26,220	3.6
Prostate	210,501	13.5			210,501	29.1
Testicular	9,768	0.6			9,768	1.3
Thyroid	2,093	0.1	1,314	0.2	779	0.1
Other cancers	520,824	33.3	217,166	25.9	303,658	41.9
Disease stage						
0	69,810	4.4	34,538	4.1	35,272	4.9
I	176,815	11.3	119,222	14.1	57,593	7.9
II	379,963	24.2	227,996	27.0	151,967	21.0
III	448,839	28.6	256,902	30.4	191,937	26.5
IV	356,699	22.7	144,805	17.1	211,894	29.2
Missing information	138,185	8.8	62,089	7.3	76,096	10.5
Treatment received						
Chemotherapy	655,013	42.7	351,650	41.2	303,363	42.3
With surgery	68,206	3.9	40,166	3.9	28,040	4.7
Radiotherapy	384,081	24.1	189,650	22.2	194,431	27.1
With surgery	35,237	2.0	17,250	2.5	17,987	2.0
Chemo- and radiotherapy	531,217	33.2	311,340	36.5	219,877	30.6
With surgery	61,486	3.5	41,542	2.8	19,944	4.8

* Race/color total available in APAC database only from 2008 onward



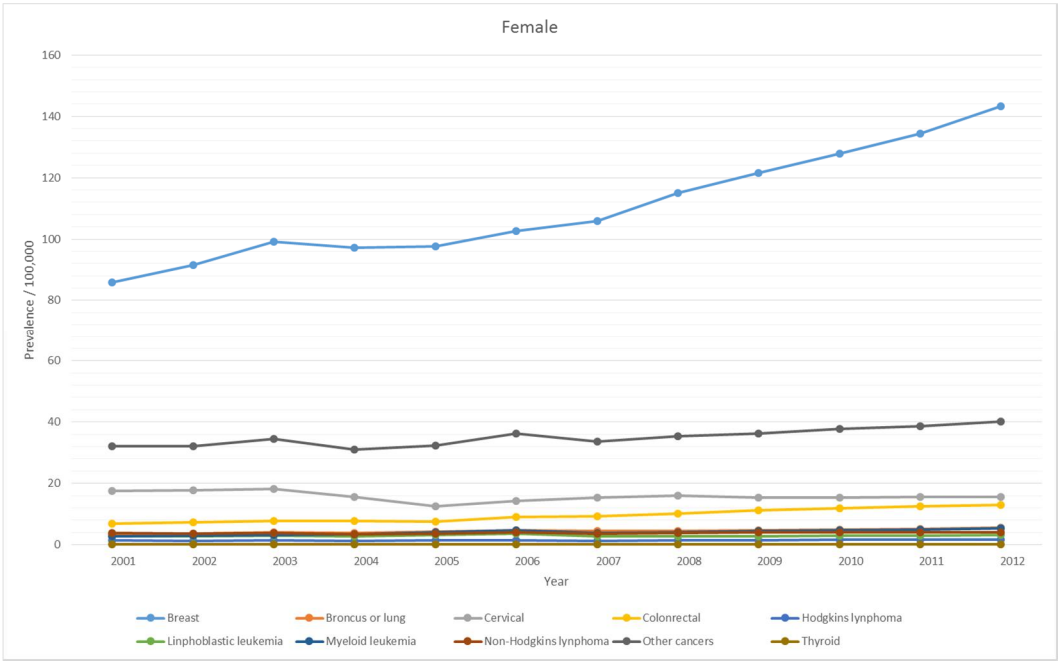
A



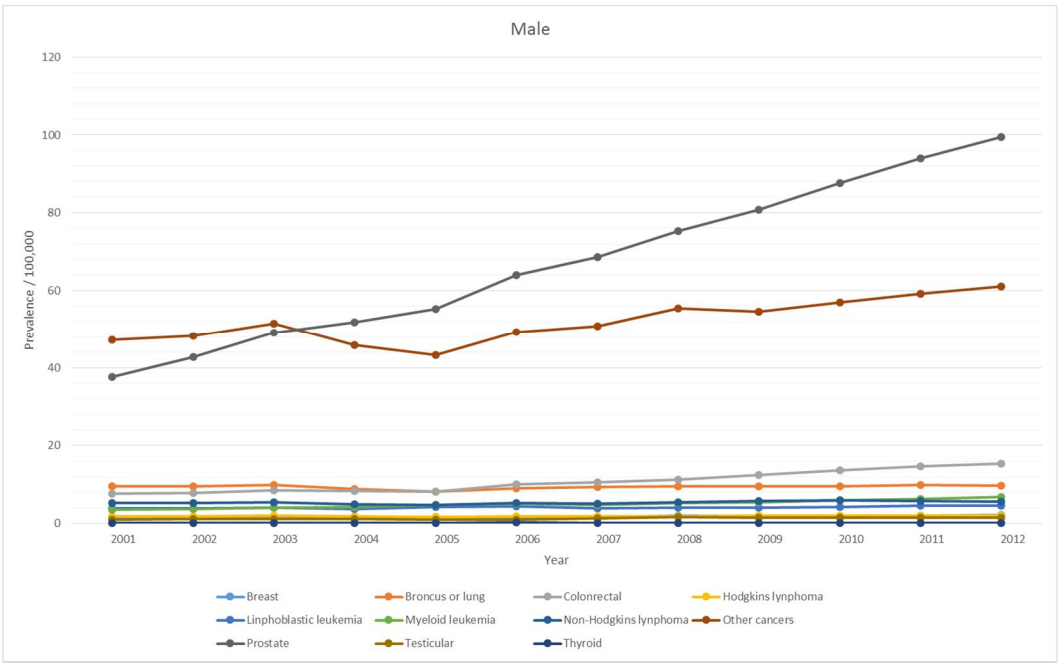
B

FIGURE 1 – Rate (/100,000 population) of initial chemo- and/or radiotherapy of cancer financed by the SUS, by sex. Panel A – males, Panel B – females, Brazil, 2001-2012. Data are age-standardized to the WHO 2000-2025 world population[15].

A



B



Supplementary FIGURE 1 – Rate (/100,000 population) of initial or ongoing chemo- and/or radiotherapy for cancer within the SUS, by sex. Panel A – males, Panel B – females, Brazil, 2001-2012. Data are age-standardized to the WHO 2000-2025 world population[15].

TABLE 3 – Temporal trends in initial chemo- and/or radiotherapy of patients with potentially curable cancers financed by the SUS, by sex. Brazil, 2001-2012.

Cancer	Female					Male					Total				
	N	Year		APC	95% CI	N	Year		APC	95% CI	N	Year		APC	95% CI
		Initial	Final				Initial	Final				Initial	Final		
Breast	389,462	2001	2012	3.3*	(1.7 ; 4.9)	53,446	2001	2012	1.8*	(0.6 ; 3)	442,908	2001	2012	2.4	(-0.3 ; 5.2)
Bronchus or lung	35,027	2001	2012	3.0*	(1.1 ; 4.9)	59,935	2001	2012	-0.3	(-1.5 ; 1.1)	94,962	2001	2012	0.8	(-0.4 ; 2.1)
Cervical	127,899	2001	2012	-1.9*	(-3.8 ; -0.0)	-	-	-	-	-	127,899	2001	2012	-1.9*	(-3.8 ; -0.0)
Colorectal	65,613	2001	2012	4.9*	(3.2 ; 6.6)	61,708	2001	2012	5.9*	(4.3 ; 7.5)	127,321	2001	2012	5.4*	(3.8 ; 7.0)
Hodgkin's lymphoma	11,123	2001	2012	1.7	(-1.0 ; 4.4)	13,064	2001	2012	0.7	(-0.8 ; 2.3)	24,187	2001	2012	1.2	(-0.8 ; 3.1)
Lymphoblastic leukemia	17,024	2001	2012	0.2	(-4.0 ; 4.7)	20,923	2001	2012	1.1	(-1.5 ; 3.7)	37,947	2001	2012	0.7	(-2.5 ; 4.1)
Myeloid leukemia	21,806	2001	2012	0.6	(-4.1 ; 5.6)	21,589	2001	2012	1.6	(-1.2 ; 4.5)	43,395	2001	2012	1.2	(-2.5 ; 4.9)
Non-Hodgkin's lymphoma	26,076	2001	2012	0.6	(-1.9 ; 3.1)	31,208	2001	2012	0.9	(-0.9 ; 2.6)	57,284	2001	2012	0.7	(-1.3 ; 2.8)
Prostate	-	-	-	-	-	233,092	2001	2012	7.0*	(5.3 ; 8.7)	233,092	2001	2012	7.0*	(5.3 ; 8.7)
Testicular	-	-	-	-	-	10,755	2001	2005	-4.5	(-11.6 ; 3.0)	10,755	2001	2005	-4.5	(-11.6 ; 3.0)
							2005	2008	22.7	(-1.0 ; 52.0)		2005	2008	22.7	(-1.0 ; 52.0)
							2008	2012	-6.0*	(-11.5 ; -0.1)		2008	2012	-6.0*	(-11.5 ; -0.1)
Thyroid	1,634	2001	2008	-4.5*	(-7.0 ; -1.9)	1,019	2001	2012	-1.6	(-4.8 ; 1.7)	2,653	2001	2012	-1.5	(-3.4 ; 0.4)
		2008	2012	5.0	(-1.2 ; 11.6)					-					-
Other cancers	256,673	2001	2012	1.1	(-0.3 ; 2.5)	334,186	2001	2012	1.6*	(0.2 ; 3.0)	590,859	2001	2012	1.3*	(0.1 ; 2.6)

Abbreviations: APC, annual percentage change (APC); CI, 95% confidence interval; '-', not applicable.

Rates are per 100,000 population and age-adjusted to the WHO 2000-2025 world population [15].

*Statistically significant (p-value < 0.05).

Table 3 shows the overall age-standardized rates of change in initiating treatment by cancer site over period 2001 through 2012, together with results of joinpoint analysis, separately for males and females of all ages. We observed a significant annual increase in initial treatments among women for breast cancer [APC=3.3% (95% CI 1.7%-4.9%)], bronchus and lung cancer [APC= 3.0% (1.1%-4.9%)], and colorectal cancer [APC= 4.9% (3.2%-6.6%)]. Thyroid cancer, in woman, shows a significant decrease in initiating treatment [APC= -4.5% (–7.0%- –1.9%)] between 2001 and 2008, followed by a non-significant annual increase of 5.0% (–1.2% - 11.6%) over the period of 2008 through 2012. In addition, cervical cancer shows a significant decrease [APC= –4.9% (–3.8%- –0.0%)] from 2001 through 2012.

Among males, we detected a significant increasing in the initial treatment of prostate [APC= 7.0% (5.3%-8.7%)], colorectal [APC= 5.9% (4.3%-7.5%)], and breast [APC= 1.8% (0.6%-3.0%)] cancer over the same period. Testicular cancer showed a decrease (APC=-4.5%) from 2001 through 2005, then a non-significant increase and finally another significant decrease from 2008-2012 [APC=-6.0 (-11.5- -0.1)], although the trend over the whole period was upward [APC= 4.1% (0.8%-7.6%)] and significant.

The average time from diagnosis to the start of treatment and the percentage of patients treated within 60 days after diagnosis, from 2008 to 2012, shows considerable variability among the subset of potentially treatable and/or frequent cancers (Table 4). During the period from 2008 to 2012, for all cancers, we observed that 62.7% (62.5; 62.9%) of patients on cancer treatment began treatment within 60 days, being 65.4% in 2008 and 62.3% in 2012.

Although at least 80% of patients started treatment in less than 60 days for Hodgkin's lymphoma, non-Hodgkin's lymphoma, lymphoblastic leukemia, myeloid leukemia, and bronchus or lung cancer, under 60% of patients started treatment within 60 days for prostate, cervical and breast cancers. Between 60% and 80% of patients with testicular, thyroid, colorectal and all other cancers

as a group, had waits of >60 days. All individual cancers evaluated present a stable tendency from 2008 to 2012 in the percentage of patients treated within 60 days after diagnosis. Additionally, this distribution of waiting time for all types of cancers is asymmetric (data not shown), with a significant portion of patients have longer waiting times than expected from a normal distribution. The Brazilian regions showed significant differences in median waiting time and percentage of patients treated within 60 days (p-value <0.01) for all cancers except thyroid (p <0.033). Overall, median time between diagnosis and initiation of treatment (percentage treated within 60 days, [CI 95%]) was 49 days (56.9%, [54.3%; 59.5%]) in the North region, 49 days (58.6%, [58.3%, 58.9%]) in the Southeast, 46 days (62.1%, [61.2%; 62.9%]) in the Central-West, 43 days (63.5%, [63.1%; 63.9%]) in the Northeast, and 34 (68.9%, [68.5%, 69.2%]) in the South.

TABLE 4 – Descriptive statistics of time (days) from diagnosis to initial treatment for patients with potentially curable cancers financed by the SUS, 2008-2012. Brazil.

Cancer	Year	N	Q1	Median	Q3	Mean (CI95%)		% patients treated up to 60 days after diagnosis (CI95%)	
Breast	2008	6,872	15	41	95.5	66.9	(65.2-68.6)	61.7	(60.6-62.9)
	2009	6,966	21	48	97	70.5	(68.9-72.2)	58.7	(57.6-59.9)
	2010	8,184	22	48	95	68.8	(67.3-70.2)	59.2	(58.1-60.2)
	2011	8,205	25	50	96	69.9	(68.5-71.3)	57.5	(56.5-58.6)
	2012	10,052	22	47	90	65.8	(64.6-67.1)	60.6	(59.6-61.6)
Bronchus or lung	2008	2,405	12	26	47	38.6	(36.8-40.4)	82.9	(81.4-84.4)
	2009	3,029	13	28	53	40.3	(38.8-41.8)	79.7	(78.3-81.2)
	2010	3,323	14	29	54	40.9	(39.5-42.4)	78.9	(77.5-80.3)
	2011	3,679	15	31	55	42.5	(41.1-43.9)	79.0	(77.7-80.3)
	2012	3,983	14	30	55	41.3	(40.0-42.6)	79.0	(77.7-80.2)
Cervical	2008	3,343	23	46	87	61.1	(59.3-62.9)	61.2	(59.6-62.9)
	2009	3,494	25	51	88	63.6	(61.8-65.3)	57.8	(56.1-59.4)
	2010	3,750	31	56	94	68.8	(67.1-70.5)	53.2	(51.6-54.8)
	2011	4,074	30	55	90	67.1	(65.5-68.7)	54.9	(53.4-56.4)
	2012	4,192	31	56	91	67.4	(65.8-68.9)	53.5	(52.0-55.0)
Colorectal	2008	2,196	22	42	74	55.2	(52.8-56.9)	65.9	(63.9-67.8)
	2009	2,632	22	42	72	53.7	(51.9-55.5)	66.1	(64.3-67.9)
	2010	3,200	24	46	75	55.7	(53.9-57.1)	64.3	(62.7-66.0)
	2011	3,591	26	48	78	58.0	(56.2-59.4)	62.6	(61.0-64.2)
	2012	4,210	24	46	76	55.2	(54.0-56.7)	63.6	(62.1-65.0)
Hodgkin's lymphoma	2008	561	10	27	52	43.7	(39.1-48.3)	79.5	(76.2-82.9)
	2009	634	11	27	53	43.6	(39.4-47.8)	79.0	(75.8-82.2)
	2010	725	11	26	49	39.4	(35.9-42.8)	81.4	(78.5-84.2)
	2011	793	11	27	52	41.5	(38.0-44.9)	80.6	(77.8-83.3)
	2012	854	10	26	50	41.4	(37.9-44.9)	80.4	(77.8-83.1)
Lymphoblastic leukemia	2008	774	0	4	21	20.9	(17.9-24.0)	91.0	(88.9-93.0)
	2009	799	0	4	18	19.2	(16.4-22.0)	91.9	(90.0-93.8)
	2010	757	1	5	21	21.3	(18.3-24.3)	89.7	(87.5-91.9)
	2011	874	1	5	22	21.2	(18.3-24.0)	91.3	(89.4-93.2)
	2012	1,013	1	6	19	21.9	(19.2-24.6)	89.0	(87.1-91.0)
Myeloid leukemia	2008	918	0	4	19	20.5	(17.8-23.2)	90.4	(88.5-92.3)
	2009	798	0	4	15	17.1	(14.6-19.6)	93.2	(91.5-95.0)
	2010	964	0.5	5	20.5	18.9	(16.6-21.1)	92.2	(90.5-93.9)
	2011	940	1	6	24	21.1	(18.7-23.6)	90.4	(88.5-92.3)
	2012	1,167	1	6	26	20.3	(18.2-22.4)	90.4	(88.7-92.1)
Non-Hodgkin's lymphoma	2008	1,207	6	21	37	32.4	(33.1-38.1)	82.4	(80.4-84.3)
	2009	1,358	7	23	49	36.2	(36.6-41.1)	80.0	(78.1-82.0)
	2010	1,486	8	23	47	36.4	(36.6-40.9)	80.1	(78.3-82.0)
	2011	1,577	7	23	46	34.8	(35.1-39.1)	81.4	(79.7-83.2)
	2012	1,755	8	23	48	35.2	(35.8-39.5)	80.3	(78.6-82.0)
Prostate	2008	4,659	29	61	113	78.0	(76.1-79.9)	48.9	(47.4-50.3)
	2009	4,496	32	68	117	82.8	(80.8-84.7)	45.0	(43.6-46.5)
	2010	5,247	36	71	122	86.0	(84.2-87.8)	42.5	(41.2-43.8)
	2011	5,801	38	75	123	87.9	(86.2-89.6)	40.1	(38.8-41.4)
	2012	7,228	35	69	119	83.4	(81.8-84.9)	43.3	(42.2-44.5)
Testicular	2008	177	10	30	54	40.6	(34.5-46.8)	79.1	(73.0-85.1)
	2009	231	10	30	61	40.5	(35.0-46.0)	73.6	(67.9-79.3)
	2010	247	8	34	73	45.8	(40.1-51.5)	68.0	(62.2-73.9)
	2011	256	8	30	58	40.8	(35.3-46.3)	77.7	(72.6-82.9)
	2012	278	10	35	69	47.3	(41.1-53.4)	69.4	(64.0-74.9)
Thyroid	2008	26	19	38	66	50.4	(32.6-68.2)	69.2	(50.2-88.2)
	2009	40	15.5	30.5	51	44.7	(28.6-60.8)	77.5	(64.0-91.0)
	2010	50	12	37	77	48.7	(34.6-62.9)	70.0	(56.8-83.2)
	2011	56	22.5	42.5	110.5	67.1	(50.3-83.8)	64.3	(51.3-77.2)
	2012	60	24	50.5	96.5	64.1	(50.0-78.1)	61.7	(49.0-74.3)
Other cancers	2008	9,134	20	42	76	55.3	(52.6-54.3)	66.8	(66.0-67.6)
	2009	10,609	21	45	81	57.7	(55.7-57.3)	63.9	(63.1-64.6)
	2010	12,010	21	45	81	57.9	(56.0-57.6)	63.4	(62.6-64.1)
	2011	12,919	23	46	80	58.8	(56.6-58.0)	62.9	(62.2-63.6)
	2012	15,356	22	46	78	58.0	(56.4-57.8)	63.6	(62.9-64.2)

DISCUSSION

This study describes, in comprehensive form for the first time the treatment of cancer with an emphasis on cancers deemed treatable and/or frequent, within the Brazilian national health system from 2001 through 2012. Among these, breast cancer, cervical cancer, colorectal cancer, and bronchus or lung cancer are the focus of most treatment for woman. For men, prostate cancer, colorectal cancer, and bronchus and lung cancer were the focus. A major expansion in treatment occurred from 2001 to 2012. In the latter period of this expansion, patient waiting time from diagnosis to the initiation of treatment was, on average, 59 days, with 62.7% initiating treatment within the congressionally mandated 60 days. This waiting time was basically unchanged over the 2008-2012 period.

While we found an increasing number of initial and overall treatment rates (Figure 1 and Supplementary Figure 1), this increase does not appear to follow the same pattern for chemo- and radiotherapy, the expansion of radiotherapy being less than for chemotherapy. The Ministry of Health´s data of deflated investment between 2000 and 2010, show that real spending in radiotherapy decreased -1.7% per year while that in chemotherapy increased 4.7% per year [19].

Brazil presents significant variability in cultural, social and economic development throughout its large territorial dimension, impacting on the occurrence and treatment of cancer[2,20]. The comparison of distribution of patients treated on the basis of APAC data to the Brazilian population distribution between 2001 through 2012 reveals that, as a group, the North, Northeast and Central West regions had 12.8% fewer chemo- and radiotherapy treatments than would be expected on the basis of the size of their populations – suggesting heterogeneity of cancer service coverage and migration for treatment, although this assertion presumes an equal or greater cancer incidence of cancers requiring treatment in these regions. Additionally, the distribution of

race/color among patients treated for chemo-radiotherapy (Table 1) showed a very different pattern compared to the same distribution presented in the Brazilian census of 2000 and 2010. According to the IBGE census [21,22], in 2000, the percentage of brown skin color (mixed ancestry) was 38.5%, and 43.1% in 2010 – a significantly greater fraction when compared to the percentage of 28.9% being of brown skin color observed in the treatment sample. Similarly, the proportion of blacks in the population in 2000 and 2010 were 6.2%, and 7.6% respectively, as opposed to only 5.5% among those treated. In contrast, the self-declared white population in 2000 and 2010 was 53.7% and 47.7% respectively – significantly below the 63.7% who received initial treatment for cancer based on APAC records.

New cases of cancer usually initiated treatment in advanced stages of their disease – 51.3% presenting stage III or IV tumors. Studies consider that the earlier treatment begins, the greater are the chances of cure [23], as delay in diagnosis is related to later stages of disease and greater risk of metastasis, and metastasis, in turn, is directly related to reduction of survival time and worse prognosis [1,24,25]. Yet overall delay in receiving treatment appears to be much more related to patient delay in seeking care and physician delay in making a diagnosis than to delay in treatment initiation once diagnosis is made. Nevertheless, it appears logical that delay in treatment initiation is not beneficial to the patient.

The trend analysis of age-standardized rates of treatment incidence in SUS shows significant increases in initial treatment, among women, for breast, bronchus or lung, and colorectal cancers; and among men, for prostate, breast, and colorectal cancer. On the other hand, we found a significant decrease in the initial treatment of cervical and thyroid cancer among women, and for men, for testicular cancer in more recent years of the period.

These increases in general are greater than those of the age-standardized incidence of cancers for Brazil over the period 2000 - 2010 as estimated by the GDB 2015 [29,30]. GBD estimates are based primarily on data from cancer registries in Brazil.

Estimates for age-standardized incidence rates of breast cancer among women in Brazil were 49.2 per 100,000 in 2000 and 63.8 per 100,000 in 2010, implying a 22.8% variation over the period. Applied APCs for rates of age-standardized incidence of treatment, estimated in SUS for the same period, results in a 38.4% variation. Similarly, for colorectal cancer we found a cumulative increase in initial treatment of 69.2, being 61.3% for women and 77.4% for men over the period. On the other hand, during the same period, GBD 2015 data revealed a cumulative increase in age-standardized incidence of 35.2% for both sexes; 33.5% among women; and 36.9% among men. We found a cumulative increase of 96.7% in initial prostate cancer treatment, in contrast to 45.7% increase estimated by GBD 2015 for its age-standardized incidence. Comparison between initial treatment and incidence for bronchus and lung cancer is somewhat compromised, as GBD 2015 analyses include tracheal cancer within their lung cancer group. Despite this methodological difference, among women, for the same period, rates of age-standardized initial treatment increased by 34.4% in SUS, vs. a 29.1% increase in incidence in GBD analyses.

In the group of cancers that showed a decreasing trend in age-standardized initial treatment, thyroid cancer presented a cumulative variation of -55.3% in SUS between 2000 through 2010, whereas GBD suggests a cumulative variation of only -25.4% in incidence. Initial treatment of cervical cancer in the SUS decreased 20.7%, while its GBD-estimated incidence decreased 34.5%.

Thus, except for thyroid cancer, expansion of initial treatment within the SUS was always greater than that expected based in changes in incidence.

According to Brazilian constitution, health is a fundamental right of the citizen, and the State is responsible for its maintenance, to be obtained by, among other things, guaranteeing the right to access to health services and to actions of health promotion. To guarantee the right to oncological health services, considering the magnitude and importance of cancer for public health, the Brazilian government passed and regulated public law 12732/12 (valid since 05/23/2013), which establishes that the first cancer treatment in the SUS should begin within a maximum period of 60 days from the signature of the pathological report, or in a shorter period according to the therapeutic need of the patient's case as registered in the medical record. Of note, no regulations were made considering earlier steps in the process, for example, the referral process[34]. In this regard, we found no evidence of trends in the time from diagnosis to treatment for the frequent and/or potentially treatable cancers selected for the study, and conclude that there was stability in this rate over the years studied.

Overall estimates, considering all cancers, for the patient's waiting time between diagnosis and beginning of treatment also varied greatly over the past few years. In 2011, the Brazilian Court of Auditors (TCU) [35], estimated that SUS patients waited, on average, up to 70 days to start chemotherapy and, on average, over 100 days to begin radiotherapy.

The Health Secretariat of the Federal District showed that, according to data from the Cancer Information System (SISCAN) in 2014, 78% of patients with early stage cancer and 79% with advanced cancer (stages III and IV) received treatment within 60 days. The Minister of Health presented, at the Fórum Estadão Saúde 2016, data of SISCAN in 2016 showing that 57% of cancer patients started treatment within 60 days [36,37].

In this study, we find results that are much closer to these latter estimates of SISCAN in 2016, especially if we assume the non-significant tendency toward a lower fraction of patients (from

65.4% to 62.3%) receiving initial treatment over the period we studied. The discrepancy of these results in relation to the other studies may be due to geographical coverage of samples used to compute estimates. Corroborating with this argument, the distribution of treatments by region of the Brazil (Table 1) suggests that the quality and coverage of cancer treatment is quite heterogeneous among Brazilian regions. Our findings thus appear to reinforce the findings of the TCU in 2011 [35] of a heterogeneous coverage of cancer treatment. This study pointed to the lack of equipment for radiotherapy in the North and Northeast regions and a non-homogeneous geographical distribution of service units across regions as causes of this heterogeneity.

Ferreira et al. [38] evaluated time elapsed from diagnosis to initiation of treatment of breast cancer between 2009 and 2011 among women attended in a relatively poor region in southern Ceará state. They observed that, on average the waiting time was 71.5 (38; 122.5) days for those financed through the SUS, similar to mean waiting time of 70.5 (68.9; 72.2) days in 2009, and 68.3 (67.6; 68.9) days for the period 2008 to 2012 we found. Moreover, Ferreira et al. show that cancer staging at diagnosis did not significantly influence the average waiting time. This study – requiring more investigation – showed a distribution of breast cancer stage similar to that which we found as well as a stability of time elapsed between diagnosis and initiation of treatment over time.

We observed that only 64% of patients with colorectal cancer started their first treatment within 60 days of diagnosis, and the mean waiting time to start treatment was 56 days. On the other hand, Valle et al. 2017 [39], evaluating male patients in a hospital in the city of São Paulo in the treatment of stomach and colorectal cancer, from July to December 2014, found an average time between diagnosis or suspected diagnosis and effective treatment of four months (122 days) and the median time 3 months. It is important to note that this statistic, although calculated including cases of stomach cancer, was based on a sample with over 70% of cases being colorectal cancer.

Concerning cervical cancer, for which our study revealed a significant reduction in the rate of new treatments in the SUS, we observed an average time of 66 days between diagnosis and start of treatment during the period of 2008 through 2012, with only 55.9% of the individuals diagnosed with this cancer being treated in up to 60 days. do Nascimento and Azevedo e Silva [40] found a worsening of median time between diagnosis and radiotherapy treatment over the period of 1995 to 2010: during the period of 2009 through 2010, the median time elapsed from diagnosis to initiation of treatment of cervical cancer was 64 days, and only 39.5% of patients received treatment within 60 days after diagnosis confirmation. Aggregating the results of this percentage of patients over the period of 2007 through 2010, they found that only 46.1% received treatment within 60 days after diagnostic confirmation [40].

In Canada, between to visit the specialist and the initiation of treatment, a patient waits on average 2.7 days for radiotherapy and 1.7 days for chemotherapy treatments[31]. In England, 99% of patients start treatment within 60 days of confirmation of diagnosis[32]. However, in India, authors estimated that this mean time was 97 days and 50% of patients start cancer treatment in 59 days[33].

Our findings of long waiting times within the context of the literature emphasize the importance of revisions, evaluation, and monitoring of all steps along the path from cancer without symptoms to treatment. It may well be that current spending on cancer screening programs could be used in a more cost-effective way if a fraction of it were allocated to improving the other causes of delay[26–28].

Data of APAC Oncology do not adequately estimate the prevalence and incidence of cancer in Brazil. Although the SUS has become more inclusive throughout the period between 2000 through 2012, the public health system presents a demand for cancer treatments greater than its network

of hospitals, qualified for this service, can offer [19,41]. In addition, due to the inequality of coverage for cancer treatment, the process of migration to regions with reasonable availability for the treatment of cancer often occurs [35,41]. Most importantly, with respect to the use of SUS estimates to indicate overall Brazilian reality, private health coverage has assumed an increasing share of cancer treatment over this period [42].

The quality of information derived from APAC Oncology limits our results in some cases. For example, race/color and tumor stage. Race/color presented missing information in 38.9% of the records of initial treatment and cancer stage in 8.8%, which could bias findings related to these variables. An additional limitation is that we cannot distinguish a truly initial treatment of cancer from a renewal of treatment upon relapse, for patients receiving their initial treatment prior to and not including 2000. Finally, for reasons beyond our control, data on treatments initiating in 2013 or later were not available for inclusion in this series, limiting its extension to more recent years.

However, on the other hand, the completeness of the date of diagnosis and date of initiating treatment present in our linked APAC database for the years 2008 to 2012 was superior to that in SISCAN, the current cancer monitoring system in Brazil [9]. The deficiency of this information in SISCAN has been high: in 2014, 56.6% of the cases in the database lacked at least one of this two dates, and in 2016 53% of registered cases [43,44].

A strength of our study is that the APAC Oncology database permits analysis of a consistent series of several years of all publicly-financed chemo- and radiotherapy treatments in Brazil, providing more information than that obtained from more localized sampling. Our study thus presents, for the first time, the epidemiological profile, created through record linkage of large databases of SUS, of a group of cancers potentially treatable and/or frequently treated in SUS for the period from 2001 thought 2012.

The incorporation of knowledge about the epidemiological history of patients treated in the public health system here produced, as well the exploration of other findings with the APAC Oncology database and the extension of its temporal series, can guide the SUS in implementing improvements in the efficiency of its cancer initiatives. Thus, this study provides an important path for future initiatives to provide continuous monitoring of chemo- and radiotherapy of cancer as well as of the trajectories of patients so-treated within the SUS.

CONCLUSION

Treatment of potentially curable cancers within the SUS expanded notably over the 12-year period starting in 2001. Time from diagnosis to treatment during this period of wider geographic reach and major expansion of coverage increased slightly, emphasizing the necessity of continuous monitoring of this indicator to guide the implementation of public policies aimed at improving care of patients with cancer.

Declarations

Acknowledgments

Funding/Sponsorship

This work was conducted during a scholarship supported by the International Cooperation Program CAPES/STICAMSUD at the Emory University. Financed by CAPES – Brazilian Federal Agency for Support and Evaluation of Graduate Education within the Ministry of Education of Brazil.

The funders had no role in study design, data collection, analysis, decision to publish, or preparation of the manuscript.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

IP, and BBD developed the study content and design. IP performed the record linkage, electronic matching, created the base analytic dataset, analyzed and interpreted the data, created the tables and figures, and wrote the first draft of the Introduction, Methods, and Results. IP and BBD all edited and approved the final version. BBD oversaw the research process. XX, YY, and ZZ consistently supplied critical inputs during review of the drafts. All authors contributed to interpreting the data, critically reviewed the drafts, and approved the final manuscript.

REFERENCE

1. GBD 2015 Mortality and Causes of Death Collaborators. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980-2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet Lond. Engl. 2016;388:1459–544.

2. Instituto Nacional de Câncer. Estimativa 2016: incidência de câncer no Brasil [Internet]. Rio de Janeiro: INCA; 2015. Available from: www.inca.gov.br/estimativa/2016/estimativa-2016-v11.pdf

3. World Health Organization, Public Health Agency of Canada, editors. Preventing chronic diseases: a vital investment. Geneva : [Ottawa]: World Health Organization ; Public Health Agency of Canada; 2005.

4. Beaglehole R, Bonita R, Horton R, Adams C, Alleyne G, Asaria P, et al. Priority actions for the non-communicable disease crisis. The Lancet. 2011;377:1438–47.

5. Beaglehole R, Horton R. Chronic diseases: global action must match global evidence. The Lancet. 2010;376:1619–21.

6. World Health Organization. Global action plan for the prevention and control of noncommunicable diseases: 2013-2020. [Internet]. 2013 [cited 2017 Mar 8]. Available from: http://apps.who.int/iris/bitstream/10665/94384/1/9789241506236_eng.pdf

7. WHO | Global Action Plan for the Prevention and Control of NCDs 2013-2020 [Internet]. WHO. [cited 2017 Mar 8]. Available from: http://www.who.int/nmh/events/ncd_action_plan/en/

8. Farmer P, Frenk J, Knaul FM, Shulman LN, Alleyne G, Armstrong L, et al. Expansion of cancer care and control in countries of low and middle income: a call to action. Lancet Lond. Engl. 2010;376:1186–93.

9. Prestes, Isaias V., Duncan, Bruce B. Record linkage of large databases of the Brazilian national health system to enable cohort analyses of cancer patients, 2000 to 2012. Popul. Health Metr. 2017;(preparation).

10. Sørensen HT, Lash TL. Use of administrative hospital registry data and a civil registry to measure survival and other outcomes after cancer. Clin. Epidemiol. 2011;3:1.

11. Instituto Brasileiro de Geografia e Estatística, editor. Pesquisa nacional de saúde, 2013: acesso e utilização dos serviços de saúde, acidentes e violências: Brasil, grandes regiões e unidades da Federação. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística-IBGE; 2015.

12. Sobin LH, Wittekind CH. UICC TNM classification of malignant tumours. 6'th ed. John Wiley & Sons, New York; 2002.

13. Ministério da Saúde Brasil. MANUAL DE BASES TÉCNICAS DA ONCOLOGIA – SIA/SUS - SISTEMA DE INFORMAÇÕES AMBULATORIAIS [Internet]. Brasília, DF, Brasil; 2013. Available from: http://bvsms.saude.gov.br/bvs/publicacoes/inca/manual_oncologia_14edicao.pdf

14. DA POPULAÇÃO EAEM. PROJEÇÃO DA POPULAÇÃO DO BRASIL POR SEXO E IDADE PARA O PERÍODO 1980-2050–Revisão 2004 Metodologia e Resultados.

15. National Cancer Institute. World (WHO 2000-2025) Standard - Standard Populations - SEER Datasets [Internet]. World WHO 2000-2025 Stand. [cited 2017 Jul 6]. Available from: https://seer.cancer.gov/stdpopulations/world.who.html

16. Kim HJ, Fay MP, Feuer EJ, Midthune DN. Permutation tests for joinpoint regression with applications to cancer rates. Stat. Med. 2000;19:335–51.

17. Joinpoint Regression Program - Surveillance Research Program [Internet]. [cited 2017 Mar 5]. Available from: https://surveillance.cancer.gov/joinpoint/

18. SELECTING THE NUMBER OF CHANGE-POINTS IN SEGMENTED LINE REGRESSION [Internet]. [cited 2017 Jul 14]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2737518/

19. Sandro Martins. Redes de Atenção à Saúde: Prioridades - Sandro Martins [Internet]. São Paulo, Brasil; 2015 [cited 2017 Aug 7]. Available from: https://www.slideshare.net/institutooncoguia/ppt-sandro-09h00

20. Guerra MR, Bustamante-Teixeira MT, Corrêa CSL, Abreu DMX de, Curado MP, Mooney M, et al. Magnitude and variation of the burden of cancer mortality in Brazil and Federation Units, 1990 and 2015. Rev. Bras. Epidemiol. Braz. J. Epidemiol. 2017;20Suppl 01:102–15.

21. IBGE. IBGE | Censo 2010 [Internet]. ATLAS Digit. Bras. 2017 [cited 2017 Aug 7]. Available from: http://censo2010.ibge.gov.br/

22. Portal Brasil de Notícias - EBC. Censo 2010 mostra as características da população brasileira — Portal Brasil [Internet]. Censo 2010 Most. Características Popul. Bras. 2012 [cited 2017 Aug 7]. Available from: http://www.brasil.gov.br/educacao/2012/07/censo-2010-mostra-as-diferencas-entre-caracteristicas-gerais-da-populacao-brasileira

23. GONZAGA SDFR, RIECHELMANN R, KALIKS R, DEL GIGLIO A. Análise do atraso no diagnóstico e tratamento do câncer de mama em um hospital público. Rev Assoc Med Bras. 2008;54:72–6.

24. Richards M, Westcombe A, Love S, Littlejohns P, Ramirez A. Influence of delay on survival in patients with breast cancer: a systematic review. The Lancet. 1999;353:1119–1126.

25. Cruz CSD, Tanoue LT, Matthay RA. Lung cancer: epidemiology, etiology, and prevention. Clin. Chest Med. 2011;32:605–644.

26. Hansen RP, Olesen F, Sørensen HT, Sokolowski I, Søndergaard J. Socioeconomic patient characteristics predict delay in cancer diagnosis: a Danish cohort study. BMC Health Serv. Res. 2008;8:49.

27. Neal RD, Tharmanathan P, France B, Din NU, Cotton S, Fallon-Ferguson J, et al. Is increased time to diagnosis and treatment in symptomatic cancer associated with poorer outcomes? Systematic review. Br. J. Cancer. 2015;112 Suppl 1:S92-107.

28. National Reporting and Learning Service. Delayed diagnosis of cancer: Thematic review [Internet]. 2010. Available from: http://www.nrls.npsa.nhs.uk/EasySiteWeb/getresource.axd?AssetID=69895

29. Fitzmaurice C, Allen C, Barber RM, Barregard L, Bhutta ZA, Brenner H, et al. Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life-years for 32 Cancer Groups, 1990 to 2015: A Systematic Analysis for the Global Burden of Disease Study. JAMA Oncol. 2017;3:524–48.

30. Institute for Health Metrics and Evaluation. GBD Results Tool | GHDx [Internet]. 2017 [cited 2017 Aug 7]. Available from: http://ghdx.healthdata.org/gbd-results-tool

31. Barua B, Rovere M, Skinner BJ. Waiting your turn. Fraser Institute; 2015.

32. NHS England. Statistics » Provider-based Cancer Waiting Times for Q1 2017-18 [Internet]. 2017 [cited 2017 Aug 11]. Available from: https://www.england.nhs.uk/statistics/statistical-work-areas/cancer-waiting-times/quarterly-prov-cwt/201718-quarterly-provider-based-cancer-waiting-times-statistics/provider-based-cancer-waiting-times-for-q1-2017-18/

33. Dwivedi AK, Dwivedi SN, Deo S, Shukla R, Pandey A, Dwivedi DK. An epidemiological study on delay in treatment initiation of cancer patients. Health (N. Y.). 2012;04:66.

34. Allotey P, Reidpath DD, Yasin S, Chan CK, de-Graft Aikins A. Rethinking health-care systems: a focus on chronicity. The Lancet. 2011;377:450–1.

35. TRIBUNAL DE CONTAS DA UNIÃO, Brasil. AUDITORIA OPERACIONAL Política Nacional de Atenção Oncológica [Internet]. TRIBUNAL DE CONTAS DA UNIÃO, Brasil; 2011. Available from: http://portal2.tcu.gov.br/portal/page/portal/TCU/comunidades/programas_governo/areas_atuacao/saude/Folder_Oncologia.pdf

36. Estadão. Número de pacientes que recorre ao SUS para tratar câncer aumenta 34% - Saúde [Internet]. Estadão. 2016 [cited 2017 Jul 6]. Available from: http://saude.estadao.com.br/noticias/geral,numero-de-pacientes-que-recorre-ao-sus-para-tratar-cancer-aumenta-34,10000068521

37. Portal Saúde, MS. Ministro da Saúde anuncia 11 prioridades para a oncologia [Internet]. Portal Saúde – Minist. Saúde – Wwwsaudegovbr. 2016 [cited 2017 Jul 6]. Available from: http://portalsaude.saude.gov.br/index.php/cidadao/principal/agencia-saude/25090-ministro-da-saude-anuncia-11-prioridades-para-a-oncologia

38. Alves Soares Ferreira N, Melo Figueiredo de Carvalho S, Engrácia Valenti V, Pinheiro Bezerra IM, Melo Teixeira Batista H, de Abreu LC, et al. Treatment delays among women with breast cancer in a low socio-economic status region in Brazil. BMC Womens Health. 2017;17:13.

39. Valle TD, Turrini RNT, Poveda V de B, Valle TD, Turrini RNT, Poveda V de B. Intervening factors for the initiation of treatment of patients with stomach and colorectal cancer. Rev. Lat. Am. Enfermagem [Internet]. 2017 [cited 2017 Aug 8];25. Available from: http://www.scielo.br/scielo.php?script=sci_abstract&pid=S0104-11692017000100333&lng=en&nrm=iso&tlng=pt

40. do Nascimento MI, Azevedo e Silva G. Waiting time for radiotherapy in women with cervical cancer. Rev. Saúde Pública [Internet]. 2016 [cited 2017 Aug 9];49. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4716650/

41. TRIBUNAL DE CONTAS DA UNIÃO. Relatório de Auditoria – Política Nacional de Atenção Oncológica [Internet]. TRIBUNAL DE CONTAS DA UNIÃO; 2011. Available from: http://portal2.tcu.gov.br/portal/page/portal/TCU/comunidades/programas_governo/areas_atuacao/saude/Oncologia%20-%20relat%C3%B3rio%20-%20vers%C3%A3o%20final.pdf

42. Brasil P. Cresce o número de cidadãos que têm algum tipo de cobertura privada [Internet]. Portal Bras. 04172012. Available from: http://www.brasil.gov.br/saude/2012/04/plano-de-saude-privado

43. Folha de São Paulo. 4 em cada 10 pacientes começam a tratar câncer só após prazo legal - 18/06/2016 - Cotidiano - Folha de S.Paulo [Internet]. 2016 [cited 2017 Aug 7]. Available from: http://www1.folha.uol.com.br/cotidiano/2016/06/1783019-4-em-cada-10-pacientes-comecam-a-tratar-cancer-so-apos-prazo-legal.shtml

44. Sandro Martins. Lei dos 60 dias, SISCAN e Portaria 140 - Sandro Martins [Internet]. São Paulo, Brasil; 2016 [cited 2017 Aug 7]. Available from: https://www.slideshare.net/institutooncoguia/lei-dos-60-dias-siscan-e-portaria-140-sandro-martins

CONCLUSÕES E CONSIDERAÇÕES FINAIS

O pareamento de dados permitiu a criação de uma base nacional de pacientes com câncer em tratamento de quimio- e/ou radioterapia no SUS, entre 2000 e 2012. A análise epidemiológica do tratamento oncológico público no Brasil mostrou que o SUS ampliou o acesso ao cuidado oncológico no âmbito de quimio e radioterapia nesse período. O tempo de espera entre diagnóstico e início do tratamento é longo e não se atenuou entre 2008 e 2012, como evidenciado pelo seu tempo médio e pelo percentual de pacientes que iniciou tratamento até 60 dias após o diagnóstico.

A metodologia desenvolvida e as informações geradas sobre o tratamento oncológico podem ser utilizadas pelo Ministério da Saúde para a construção de indicadores e para o monitoramento desse serviço no SUS, em conformidade com o Plano de Ações Estratégicas para o Enfrentamento das Doenças Crônicas Não Transmissíveis até 2025. Nesse sentido, esse trabalho ilustra a importância de o governo estabelecer parcerias com a academia para a resolução de problemas que merecem investigação. A experiência vivida me mostrou que a pesquisa colaborativa governo/academia pode atrair jovens talentos, com energia, e conhecimentos específicos, para que juntos possam críticas construtivas ao longo da parceria. Essa é uma característica das ações do Ministério da Saúde no Brasil que, acredito, tem sido bem-sucedida e deveria ser mais estimulada.

O pareamento de dados foi a ferramenta básica para tornar este trabalho possível. A utilização do pareamento de dados para subsidiar a vigilância e o monitoramento de ações de saúde cresceu muito no País e no mundo. O cruzamento de informações entre distintas bases de dados do SUS, além de responder a questões científicas inerentes ao campo acadêmico específico, amplia, o entendimento do comportamento de doenças, ações e serviços pelas autoridades governamentais.

Os experimentos para conversão de base de dados administrativa em base de dados analisável constituíram um processo longo e penoso, que consumiu meses de pesquisa para descoberta de soluções, testes e avaliação das técnicas e procedimentos, e também do processamento computacional. Brevemente,

organizamos a série histórica das APAC de químio e radioterapia e ligamos os registros dos pacientes atribuindo um identificador único ï a chave para a porta de acesso aos dados no plano epidemiológico. O identificador único do paciente, o número do cluster de registros, solucionou o problema decorrente da mudança de número e tipo do documento de identificação dos pacientes ao longo dos anos (ou seja, CPF e CNS que funcionavam como chave primária de uma mesma entidade, e casos de um indivíduo ter mais que um número de CNS). Um grande legado deste trabalho é a possibilidade da aplicação da mesma metodologia desenvolvida sobre outras bases de dados de APAC, abrindo caminho para a exploração do comportamento de outras doenças no SUS. A saber, a criação de uma coorte de pacientes em tratamento de doença cardiovascular, com base na APAC de laudos diversos, vinculada com informações de mortalidade, de hospitalizações e de outros procedimentos de alta complexidade no SUS.

Uma metodologia de pareamento de dados não é algo que encontra um estágio final de desenvolvimento. Assim como programas de computador, de procedimentos de ações humanas para atingir um fim, a estratégia de pareamento de dados pode sempre ser melhorada ao longo do tempo: a medida em que se pode melhor explorar o ponto a que se chegou, com a exploração de técnicas existentes, e com a aplicação e avaliação das técnicas que surgem frequentemente. O pareamento de dados é ainda um campo para muita exploração ï seja em sua esfera computacional, seja em sua esfera de teoria matemática.

Por fim, a importância das contribuições dadas por este trabalho ï ainda que histórica olhando de hoje ï , bem como a continuidade da exploração das grandes bases de dados do SUS para outras DCNT, se tornam maiores à medida em que a população brasileira experiência seu maior envelhecimento.

ANEXOS

a. Aprovação pelo Comitê da Ética e Pesquisa

HCPA - HOSPITAL DE CLÍNICAS DE PORTO ALEGRE
GRUPO DE PESQUISA E PÓS-GRADUAÇÃO

COMISSÃO CIENTÍFICA E COMISSÃO DE PESQUISA E ÉTICA EM SAÚDE

A Comissão Científica e a Comissão de Pesquisa e Ética em Saúde, que é reconhecida pela Comissão Nacional de Ética em Pesquisa (CONEP)/MS como Comitê de Ética em Pesquisa do HCPA e pelo Office For Human Research Protections (OHRP)/USDHHS, como Institutional Review Board (IRB00000921) analisaram o projeto:

Projeto: 100056 **Versão do Projeto:** 19/02/2010

Pesquisadores:
BRUCE BARTHOLOW DUNCAN

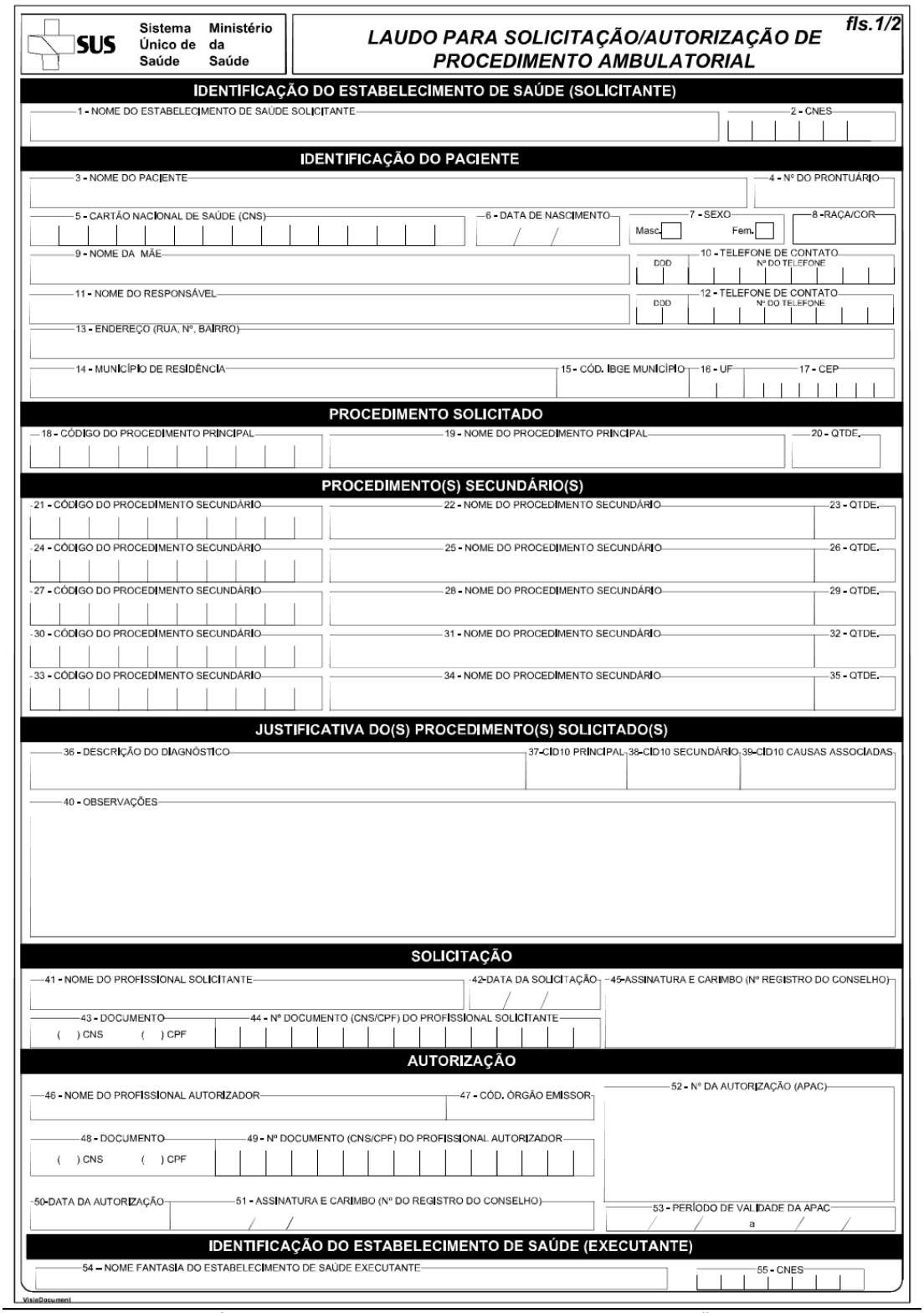
Título: Projeto de Desenvolvimento para a Consolidação do Centro Colaborador para a Vigilância do Diabetes, Doenças Cardiovasculares e Outras Doenças Não Transmissíveis - Análise de Dados Primários e Secundários dos Grandes Sistemas Nacionais de Informações em Saúde do Sistema Único de Saúde

Este projeto foi Aprovado em seus aspectos éticos e metodológicos de acordo com as Diretrizes e Normas Internacionais e Nacionais, especialmente as Resoluções 196/96 e complementares do Conselho Nacional de Saúde. Os membros do CEP/HCPA não participaram do processo de avaliação dos projetos onde constam como pesquisadores. Toda e qualquer alteração do Projeto deverá ser comunicada imediatamente ao CEP/HCPA.

Porto Alegre, 10 de março de 2010.

Profª Nadine Clausell
Coordenadora GPPG e CEP/HCPA

b. Laudo para solicitação de APAC



SUS — Sistema Único de Saúde — Ministério da Saúde

LAUDO PARA SOLICITAÇÃO/AUTORIZAÇÃO DE PROCEDIMENTO AMBULATORIAL — fls.1/2

IDENTIFICAÇÃO DO ESTABELECIMENTO DE SAÚDE (SOLICITANTE)

1 - NOME DO ESTABELECIMENTO DE SAÚDE SOLICITANTE | 2 - CNES

IDENTIFICAÇÃO DO PACIENTE

3 - NOME DO PACIENTE | 4 - Nº DO PRONTUÁRIO

5 - CARTÃO NACIONAL DE SAÚDE (CNS) | 6 - DATA DE NASCIMENTO | 7 - SEXO (Masc. / Fem.) | 8 - RAÇA/COR

9 - NOME DA MÃE | 10 - TELEFONE DE CONTATO (DDD / Nº DO TELEFONE)

11 - NOME DO RESPONSÁVEL | 12 - TELEFONE DE CONTATO (DDD / Nº DO TELEFONE)

13 - ENDEREÇO (RUA, Nº, BAIRRO)

14 - MUNICÍPIO DE RESIDÊNCIA | 15 - CÓD. IBGE MUNICÍPIO | 16 - UF | 17 - CEP

PROCEDIMENTO SOLICITADO

18 - CÓDIGO DO PROCEDIMENTO PRINCIPAL | 19 - NOME DO PROCEDIMENTO PRINCIPAL | 20 - QTDE.

PROCEDIMENTO(S) SECUNDÁRIO(S)

21 - CÓDIGO DO PROCEDIMENTO SECUNDÁRIO | 22 - NOME DO PROCEDIMENTO SECUNDÁRIO | 23 - QTDE.

24 - CÓDIGO DO PROCEDIMENTO SECUNDÁRIO | 25 - NOME DO PROCEDIMENTO SECUNDÁRIO | 26 - QTDE.

27 - CÓDIGO DO PROCEDIMENTO SECUNDÁRIO | 28 - NOME DO PROCEDIMENTO SECUNDÁRIO | 29 - QTDE.

30 - CÓDIGO DO PROCEDIMENTO SECUNDÁRIO | 31 - NOME DO PROCEDIMENTO SECUNDÁRIO | 32 - QTDE.

33 - CÓDIGO DO PROCEDIMENTO SECUNDÁRIO | 34 - NOME DO PROCEDIMENTO SECUNDÁRIO | 35 - QTDE.

JUSTIFICATIVA DO(S) PROCEDIMENTO(S) SOLICITADO(S)

36 - DESCRIÇÃO DO DIAGNÓSTICO | 37-CID10 PRINCIPAL | 38-CID10 SECUNDÁRIO | 39-CID10 CAUSAS ASSOCIADAS

40 - OBSERVAÇÕES

SOLICITAÇÃO

41 - NOME DO PROFISSIONAL SOLICITANTE | 42-DATA DA SOLICITAÇÃO | 45-ASSINATURA E CARIMBO (Nº REGISTRO DO CONSELHO)

43 - DOCUMENTO ( ) CNS ( ) CPF | 44 - Nº DOCUMENTO (CNS/CPF) DO PROFISSIONAL SOLICITANTE

AUTORIZAÇÃO

46 - NOME DO PROFISSIONAL AUTORIZADOR | 47 - CÓD. ÓRGÃO EMISSOR | 52 - Nº DA AUTORIZAÇÃO (APAC)

48 - DOCUMENTO ( ) CNS ( ) CPF | 49 - Nº DOCUMENTO (CNS/CPF) DO PROFISSIONAL AUTORIZADOR

50-DATA DA AUTORIZAÇÃO | 51 - ASSINATURA E CARIMBO (Nº DO REGISTRO DO CONSELHO) | 53 - PERÍODO DE VALIDADE DA APAC (a)

IDENTIFICAÇÃO DO ESTABELECIMENTO DE SAÚDE (EXECUTANTE)

54 – NOME FANTASIA DO ESTABELECIMENTO DE SAÚDE EXECUTANTE | 55 - CNES

VisioDocument

SUS Sistema Único de Saúde — Ministério da Saúde

LAUDO PARA SOLICITAÇÃO/AUTORIZAÇÃO DE PROCEDIMENTO AMBULATORIAL - DADOS COMPLEMENTARES

fls. 2/2

1 - ONCOLOGIA

IDENTIFICAÇÃO PATOLÓGICA DO CASO

56-Localização do tumor primário | 57-CID-10 Topografia

58-LINFONODOS REGIONAIS INVADIDOS: SIM ☐ NÃO ☐ — NÃO AVALIÁVEIS ☐ | 59-Localização de Metástase(s)

60-Estádio(UICC) | 61-Estádio (outro sistema) | 62-Grau Histopatológico

63-Diagnóstico Cito/Histopatológico | 64-Data

1.1 - QUIMIOTERAPIA

65 - TRATAMENTO(S) ANTERIOR(ES) ☐ SIM ☐ NÃO

Tratamento(s) Anterior(es)	66-Descrição	67-Data de Início
1º		
2º		
3º		

TRATAMENTO SOLICITADO - Planejamento Terapêutico Global

68-Continuidade de Tratamento ☐ NÃO ☐ SIM | 69-Data de Início do Tratamento Solicitado | 70-ESQUEMA (Sigla ou abreviatura) | 71-Nº Total de Meses Planejados | 72-Nº de Meses Autorizados

1.2 - RADIOTERAPIA

73 - TRATAMENTO(S) ANTERIOR(ES) ☐ SIM ☐ NÃO

Tratamento(s) Anterior(es)	74 - Descrição	75 - Data de Início
1º		
2º		
3º		

TRATAMENTO SOLICITADO - Planejamento Terapêutico Global

76 - Continuidade de Tratamento ☐ NÃO ☐ SIM | 77 - Data de Início do Tratamento Solicitado

78 - Finalidade: ☐ RADICAL ☐ ADJUVANTE ☐ ANTIÁLGICA ☐ PALIATIVA ☐ PRÉVIA ☐ ANTIHEMORRÁGICA

ÁREA IRRADIADA

79 - CID Topográfico		80 - Descrição	81 - Nº Campo/Incerções	82 - Data de Início	83 - Data de Término
	1				
	2				
	3				

2 – NEFROLOGIA

84-PRIMEIRO ATENDIMENTO

DATA DA 1ª DIÁLISE REALIZADA

Altura ___ m — IMC(kg/m²)
Peso ___ Kg
Diurese ___ ml
Glicose ___ mg/dl
Albumina ___ %
Hb ___ g%

ACESSO VASCULAR ☐ Sim ☐ Não
aa HIV ☐ Positivo ☐ Negativo
aa HCV ☐ Positivo ☐ Negativo
HBs Ag ☐ Positivo ☐ Negativo
Ultrasonografia Abdominal ☐ Sim ☐ Não

85-SEGUIMENTO

TRU ___
Hb ___ g%
Albumina ___ g%
Intervenção de Fístola — QTD ___

Inscrito na lista da CNCDO ☐ Sim ☐ Não
aa HIV ☐ Positivo ☐ Negativo
aa HCV ☐ Positivo ☐ Negativo
HBs Ag ☐ Positivo ☐ Negativo

SOLICITAÇÃO	AUTORIZAÇÃO
86-ASSINATURA E CARIMBO (Nº REGISTRO DO CONSELHO) PROFISSIONAL SOLICITANTE	87-ASSINATURA E CARIMBO (Nº REGISTRO DO CONSELHO) PROFISSIONAL AUTORIZADOR

c. Descrição de variáveis e tipo de dado das APAC de oncologia (2000-2007)

Arquivo de **Pacientes em Tratamento de Radioterapia** de cada estado, para cada mês de competência:

- Nomenclatura: PRufaamm.DBF (uf = sigla da Unidade da Federação, aa = ano de referência, mm = mês de referência)
- Formato: DBF
- Descrição: contém registros com informações cadastrais dos pacientes em tratamento de radioterapia
- Utilizado por: aplicativos TAB, TABX, TABWIN

Campo	*Nome*	*Descrição do campo*
1	PAR_CONDIC	CONDIÇÃO DE GESTÃO DO MUNICÍPIO:
		PB = ATENÇÃO BÁSICA
		MP = MUNICÍPIO PLENO
		MS = MUNICÍPIO SEMIPLENO
		EC = ESTADUAL CONVENCIONAL
		EA = ESTADUAL AVANÇADO
		EP = ESTADUAL PLENO
2	PAR_GESTAO	CÓDIGO DE GESTÃO DO MUNICÍPIO
3	PAR_CODUNI	CÓDIGO DA UNIDADE PRESTADORA DO SERVIÇO (SEM DÍGITO VERIFICADOR)
4	PAR_NUM	NÚMERO DA APAC
5	PAR_DATREF	DATA DE COMPETÊNCIA DO ATENDIMENTO
6	PAR_CPFPCN	CPF DO PACIENTE
7	PAR_UFNASC	UNIDADE DA FEDERAÇÃO ONDE NASCEU O PACIENTE
8	PAR_CEPPCN	CEP DO ENDEREÇO DO PACIENTE
9	PAR_MUNPCN	MUNICÍPIO DO ENDEREÇO DO PACIENTE (CÓDIGO DA UF+CÓDIGO DO MUNICÍPIO)
10	PAR_NASCPC	DATA DO NASCIMENTO DO PACIENTE

Campo	*Nome*	*Descrição do campo*
11	PAR_SEXOPC	SEXO DO PACIENTE. (M-MASCULINO; F-FEMININO)
12	PAR_INITRA	DATA DO PRIMEIRO TRATAMENTO REALIZADO
13	PAR_DIAGPR	CID (CÓDIGO INTERNACIONAL DE DOENÇAS) DA PATOLOGIA QUE ORIGINOU A DOENÇA
14	PAR_DIAGSE	CID (CÓDIGO INTERNACIONAL DE DOENÇAS) DA DOENÇA
15	PAR_METAST	INDICAÇÃO DE METASTASE (S-SIM; N-NÃO)
16	PAR_FINALI	FINALIDADE DO TRATAMENTO
17	PAR_DTDIAG	DATA (AAMM) DO DIAGNÓSTICO DO TUMOR
18	PAR_DTTRA1	DATA (AAMM) DO 1.º TRATAMENTO
19	PAR_DTTRA2	DATA (AAMM) DO 2.º TRATAMENTO
20	PAR_DTTRA3	DATA (AAMM) DO 3.º TRATAMENTO
21	PAR_ESTAD	CÓDIGO DO ESTÁDIO DO TUMOR
22	PAR_AREA1	CID TOPOGRÁFICO DA ÁREA IRRADIADA 1
23	PAR_AREA2	CID TOPOGRÁFICO DA ÁREA IRRADIADA 2
24	PAR_AREA3	CID TOPOGRÁFICO DA ÁREA IRRADIADA 3
25	PAR_CAMPOS	NÚMERO DE CAMPOS PLANEJADOS

Arquivo de **Pacientes em Tratamento de Quimioterapia** de cada estado, para cada mês de competência:

- Nomenclatura: PQufaamm.DBF (uf = sigla da Unidade da Federação, aa = ano de referência, mm = mês de referência)
- Formato: DBF
- Descrição: contém registros com informações cadastrais dos pacientes em tratamento de quimioterapia
- Utilizado por: aplicativos TAB, TABX, TABWIN

Campo	*Nome*	*Descrição do campo*
1	PAQ_CONDIC	CONDIÇÃO DE GESTÃO DO MUNICÍPIO: PB = ATENÇÃO BÁSICA MP = MUNICÍPIO PLENO MS = MUNICÍPIO SEMI-PLENO EC = ESTADUAL CONVENCIONAL EA = ESTADUAL AVANÇADO EP = ESTADUAL PLENO

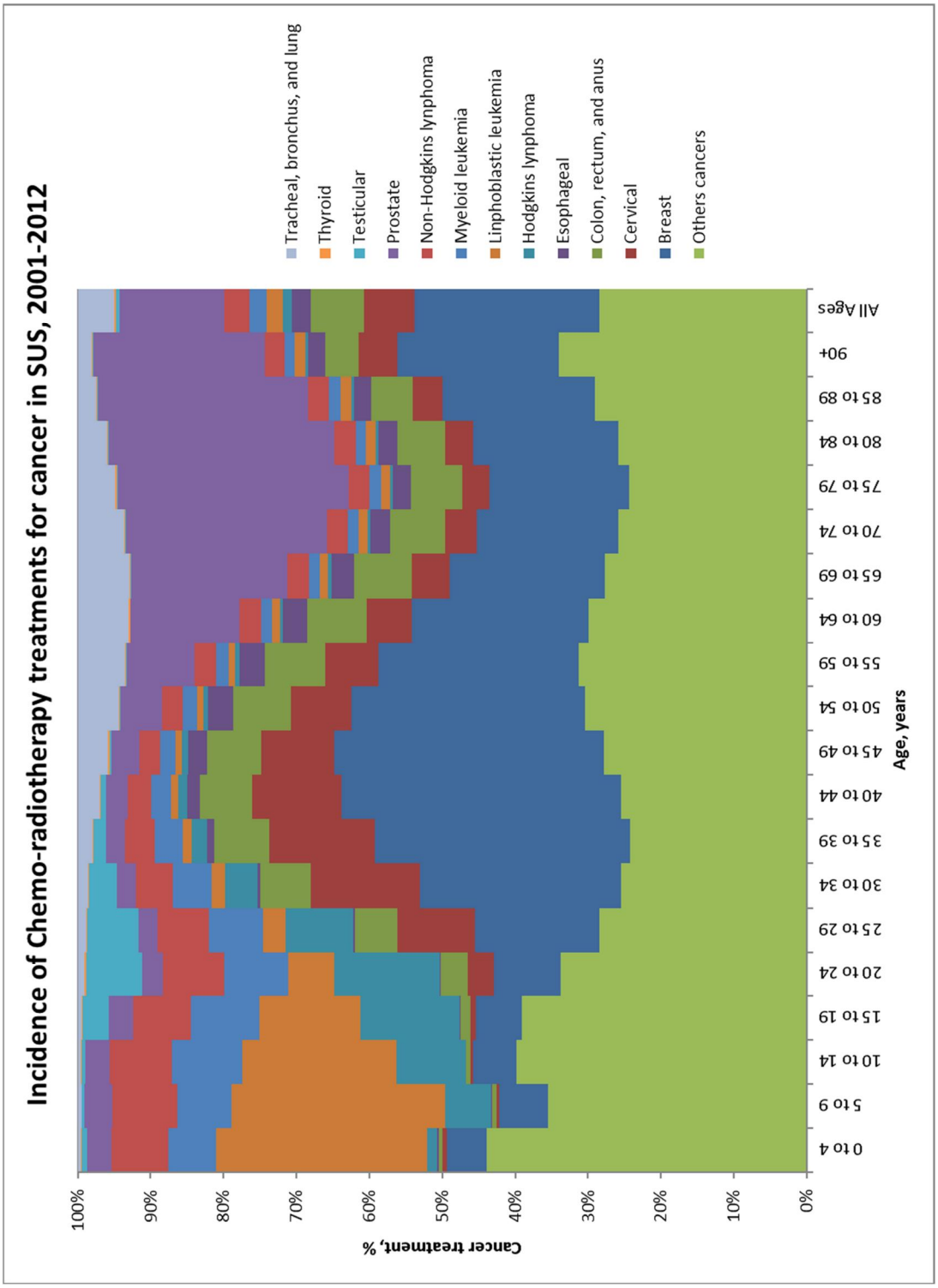
Campo	*Nome*	*Descrição do campo*
2	PAQ_GESTAO	CÓDIGO DE GESTÃO DO MUNICÍPIO
3	PAQ_CODUNI	CÓDIGO DA UNIDADE PRESTADORA DO SERVIÇO (SEM DÍGITO VERIFICADOR)
4	PAQ_NUM	NÚMERO DA APAC
5	PAQ_DATREF	DATA DE COMPETÊNCIA DO ATENDIMENTO
6	PAQ_CPFPCN	CPF DO PACIENTE
7	PAQ_UFNASC	UNIDADE DA FEDERAÇÃO ONDE NASCEU O PACIENTE
8	PAQ_CEPPCN	CEP DO ENDEREÇO DO PACIENTE
9	PAQ_MUNPCN	MUNICÍPIO DO ENDEREÇO DO PACIENTE (CÓDIGO DA UF+CÓDIGO DO MUNICÍPIO)
10	PAQ_NASCPC	DATA DO NASCIMENTO DO PACIENTE
11	PAQ_SEXOPC	SEXO DO PACIENTE. (M-MASCULINO; F-FEMININO)
12	PAQ_INITRA	DATA DO PRIMEIRO TRATAMENTO REALIZADO
13	PAQ_DIAGPR	CID (CÓDIGO INTERNACIONAL DE DOENÇAS) DA PATOLOGIA QUE ORIGINOU A DOENÇA
14	PAQ_DIAGSE	CID (CÓDIGO INTERNACIONAL DE DOENÇAS) DA DOENÇA
15	PAQ_METAST	INDICAÇÃO DE METASTASE (S-SIM; N-NÃO)
16	PAQ_DTDIAG	DATA (AAMM) DO DIAGNÓSTICO DO TUMOR
17	PAQ_DTTRA1	DATA (AAMM) DO 1.º TRATAMENTO
18	PAQ_DTTRA2	DATA (AAMM) DO 2.º TRATAMENTO
19	PAQ_DTTRA3	DATA (AAMM) DO 3.º TRATAMENTO
20	PAQ_ESTAD	CÓDIGO DO ESTÁDIO DO TUMOR
21	PAQ_MEPREV	NÚMERO DE CAMPOS PLANEJADOS

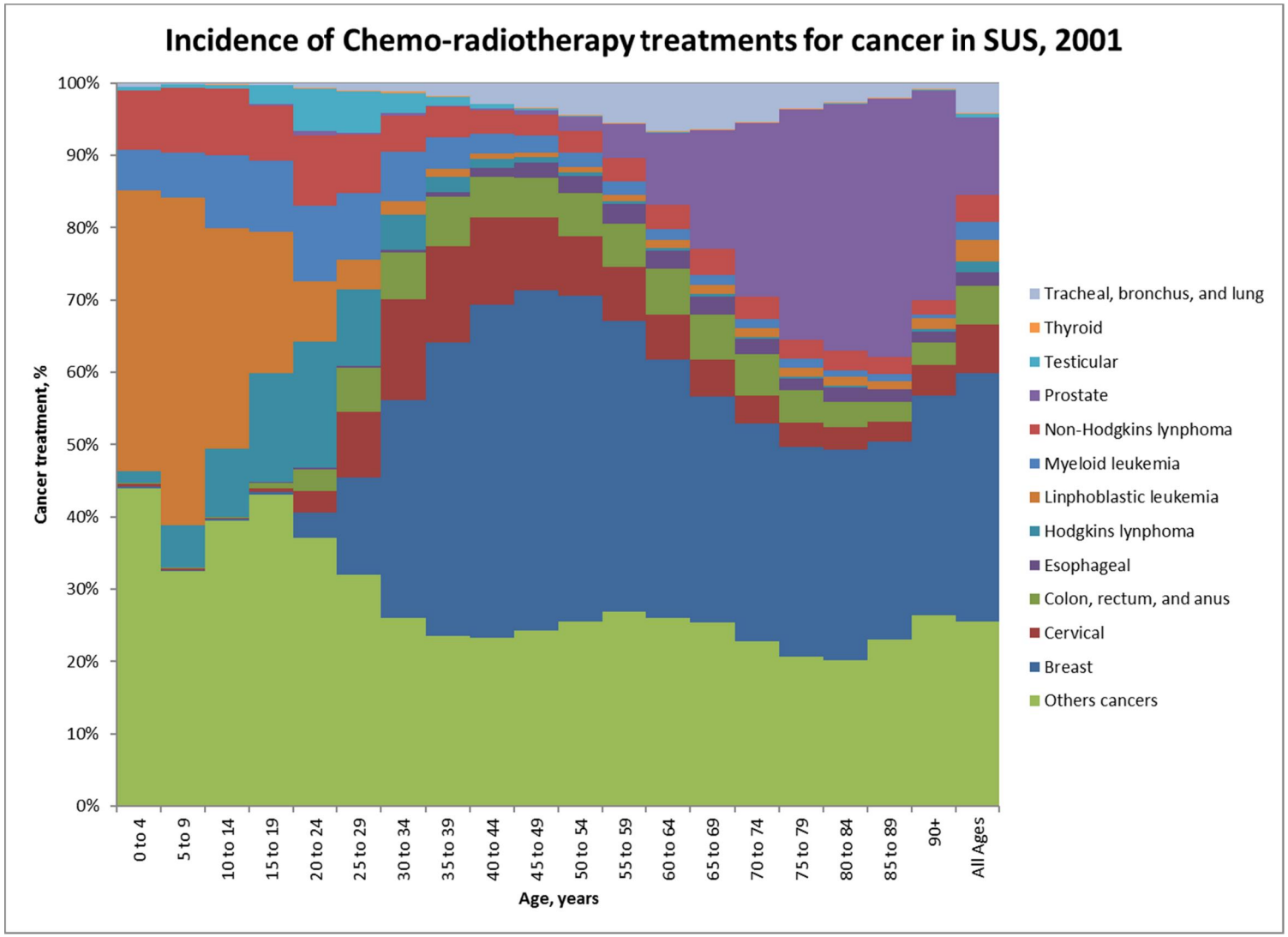
- Nomenclatura: ACufaamm.DBF (uf = sigla da Unidade da Federação, aa = ano de referência, mm = mês de referência)
- Formato: DBF
- Descrição: contém registros com informações cadastrais dos pacientes em Terapia Renal Substitutiva
- Utilizado por: aplicativos TAB, TABX, TABWIN

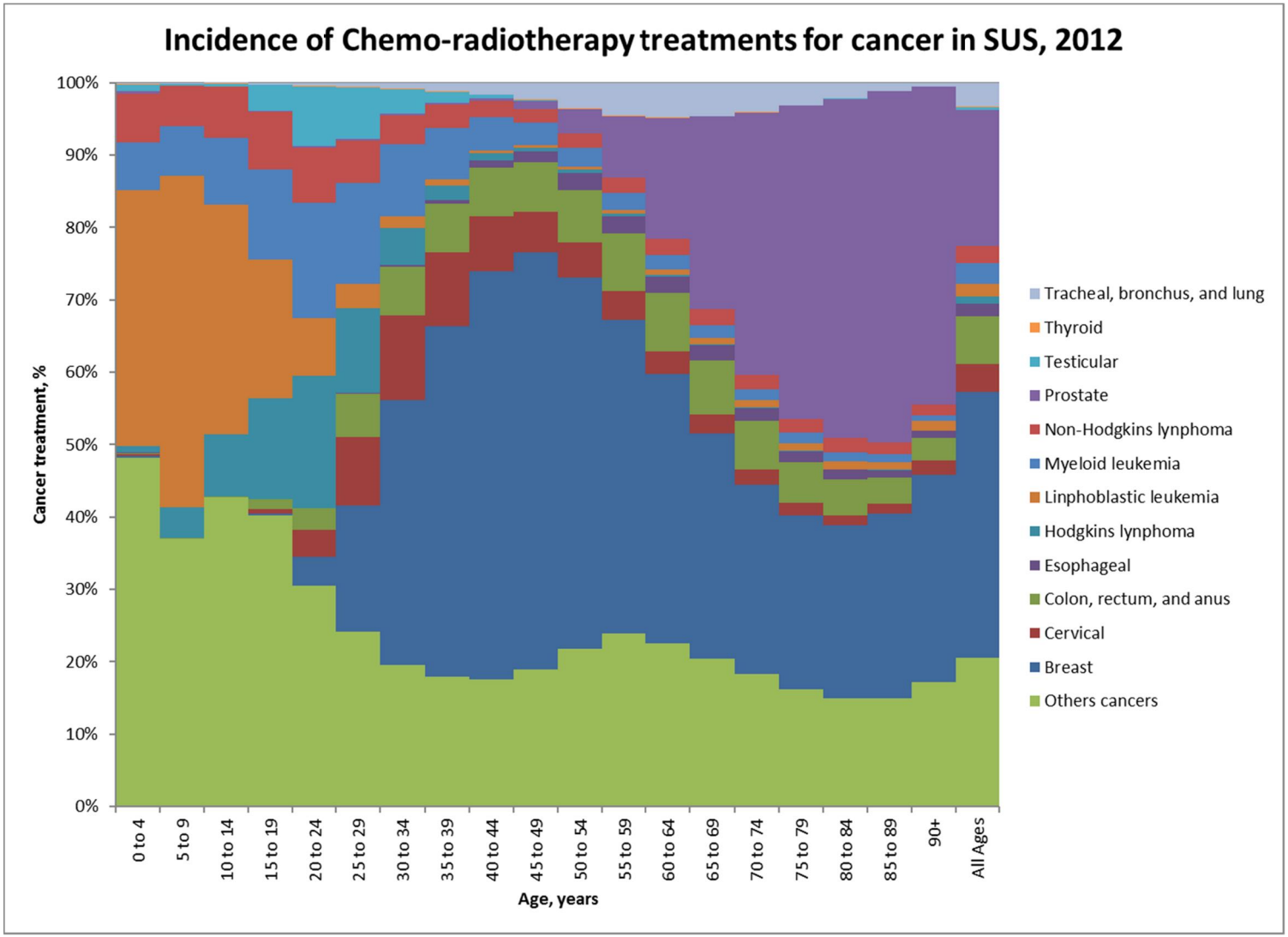
Descrição dos registros

Campo	*Nome*	*Descrição do campo*
1	APA_CONDIC	CONDIÇÃO DE GESTÃO DO MUNICÍPIO: PB = ATENÇÃO BÁSICA MP = MUNICÍPIO PLENO MS = MUNICÍPIO SEMIPLENO EC = ESTADUAL CONVENCIONAL EA = ESTADUAL AVANÇADO EP = ESTADUAL PLENO
2	APA_GESTAO	CÓDIGO DE GESTÃO DO MUNICÍPIO
3	APA_CODUNI	CÓDIGO DA UNIDADE PRESTADORA DO SERVIÇO (SEM DÍGITO VERIFICADOR)
4	APA_NUM	NÚMERO DA APAC
5	APA_DATREF	DATA DE COMPETÊNCIA DO ATENDIMENTO
6	APA_DATEM	DATA DE EMISSÃO DA APAC
7	APA_DTINIV	DATA DE INÍCIO DE VALIDADE DA APAC
8	APA_DTFIM	DATA DE FIM DE VALIDADE DA APAC
9	APA_TIPATE	TIPO DE ATENDIMENTO COBRADO POR APAC. (13-TRS; 14-RADIOTERAPIA; 15-QUIMIOTERAPIA)
10	APA_TIPAPA	TIPO DE APAC EMITIDA. (1-INICIAL; 2-CONTINUAÇÃO)
11	APA_CPFPCN	CPF DO PACIENTE
12	APA_CPFRES	CPF DO MÉDICO RESPONSÁVEL PELO ATENDIMENTO
13	APA_NOMERE	NOME DO MÉDICO RESPONSÁVEL PELO ATENDIMENTO
14	APA_PRIPAL	CÓDIGO DO PROCEDIMENTO PRINCIPAL COBRADO PELO ATENDIMENTO
15	APA_MOTCOB	CÓDIGO DO MOTIVO DA COBRANÇA DO PROCEDIMENTO
16	APA_DTOCOR	DATA DA OCORRÊNCIA NO CASO DE ALTA ÓBITO, MUDANÇA DE PROCEDIMENTO OU TRANSFERÊNCIA
17	APA_CPFDIR	CPF DO DIRETOR DA UNIDADE PRESTADORA DO ATENDIMENTO

d. Outros aspectos metodológicos adicionais




Incidence of Chemo-radiotherapy treatments for cancer in SUS, 2001
Cancer treatment, %
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0 to 4
5 to 9
10 to 14
15 to 19
20 to 24
25 to 29
30 to 34
35 to 39
40 to 44
45 to 49
50 to 54
55 to 59
60 to 64
65 to 69
70 to 74
75 to 79
80 to 84
85 to 89
90+
All Ages
Age, years
Tracheal, bronchus, and lung
Thyroid
Testicular
Prostate
Non-Hodgkins lynphoma
Myeloid leukemia
Linphoblastic leukemia
Hodgkins lynphoma
Esophageal
Colon, rectum, and anus
Cervical
Breast
Others cancers


Incidence of Chemo-radiotherapy treatments for cancer in SUS, 2012
Cancer treatment, %
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0 to 4
5 to 9
10 to 14
15 to 19
20 to 24
25 to 29
30 to 34
35 to 39
40 to 44
45 to 49
50 to 54
55 to 59
60 to 64
65 to 69
70 to 74
75 to 79
80 to 84
85 to 89
90+
All Ages
Age, years
Tracheal, bronchus, and lung
Thyroid
Testicular
Prostate
Non-Hodgkins lynphoma
Myeloid leukemia
Linphoblastic leukemia
Hodgkins lynphoma
Esophageal
Colon, rectum, and anus
Cervical
Breast
Others cancers