Ministério da Saúde
FIOCRUZ
Fundação Oswaldo Cruz

ICICT
Instituto de Comunicação e Informação
Científica e Tecnológica em Saúde

Raphael de Freitas Saldanha

Da aquisição a visualização de dados: aplicações da ciência de dados em saúde

Rio de Janeiro
2021

Raphael de Freitas Saldanha

Da aquisição a visualização de dados: aplicações da ciência de dados em saúde

Tese apresentada ao Programa de Pós-Graduação em Informação e Comunicação em Saúde do Instituto de Comunicação e Informação em Saúde (Icict) para obtenção do grau de Doutor em Ciências.

Orientador: Dr. Christovam Barcellos
Co-orientador: Dr. Marcel de Moraes Pedroso

Rio de Janeiro
2021

de Freitas Saldanha, Raphael.

Da aquisição a visualização de dados: aplicações da ciência de dados em saúde / Raphael de Freitas Saldanha. - Rio de janeiro, 2021.
167 f.; il.

Tese (Doutorado) – Instituto de Comunicação e Informação Científica e Tecnológica em Saúde, Pós-Graduação em Informação e Comunicação em Saúde, 2021.

Orientador: Christovam Barcellos.
Co-orientador: Marcel de Moraes Pedroso.

Bibliografia: f. 156-164

1. Ciência de Dados. 2. Métodos epidemiológicos. 3. Saúde pública. I. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da Biblioteca de Manguinhos/ICICT com os dados fornecidos pelo(a) autor(a).

Fundação Oswaldo Cruz
Instituto de Comunicação e Informação Científica e Tecnológica em Saúde
Programa de Pós-Graduação em Informação e Comunicação em Saúde
Raphael de Freitas Saldanha

Da aquisição a visualização de dados: aplicações da ciência de dados em saúde

Aprovado em 23 de fevereiro de 2021.

Dr. Christovam Barcellos
Orientador

Dr. Marcel de Moraes Pedroso
Co-Orientador

Dr. Paulo Roberto Borges de Souza Júnior

Dr. Ricardo Antunes Dantas de Oliveira

Dr. Fábio André Machado Porto

Dr. Eduardo Ogasawara

Rio de Janeiro
2021

*Dedico este trabalho à minha esposa Nathália, ao meu pai (*in memoriam*) e à minha mãe.*

Agradecimentos

Agradeço a minha esposa Nathália pela paciência nos momentos de aflição, pelo apoio frente às incertezas e pelo comedimento frente aos exageros, tanto na escrita desta tese quanto na vida.

Agradeço à Nathália, ao meu pai (*in memoriam*) e à minha mãe, que tanto me apoiam incondicionalmente na progressão de minha carreira acadêmica.

Agradeço aos meus professores da graduação, mestrado e doutorado. À Universidade Federal de Juiz de Fora e Fiocruz, instituições públicas, gratuitas e de qualidade.

Agradeço ao Programa de Pós-Graduação em Comunicação e Informação em Saúde, ao Instituto de Comunicação e Informação Científica e Tecnológica em Saúde, ao Laboratório de Informação em Saúde, ao Núcleo de Geoprocessamento e à Plataforma de Ciência de Dados aplicada à Saúde, pelas oportunidades e experiências.

Agradeço ao meu orientador, Prof. Christovam, e ao meu co-orientador, Prof. Marcel, pelo aconselhamento e paciência.

Study hard what interests you the most,
in the most undisciplined, irreverent and original manner possible.
(Richard Feynmann)

Resumo

Termos como *big data* e *ciência de dados* ganharam visibilidade nas últimas décadas, sendo em geral associados aos desafios de análise de grandes bases de dados. Pretende-se com este trabalho reunir teorias e métodos da Ciência de Dados em Saúde como uma contribuição à Saúde Pública, estudando o ciclo de geração e disseminação de informação em saúde e aplicar técnicas de coleta, extração e visualização de dados com métodos de *ciência de dados*. Partindo de uma perspectiva histórica, a relação entre dados e saúde foi visitada, apresentando um novo paradigma da *ciência de dados em saúde*, considerando as possibilidades hibridas de uma ciência *theory & data driven* para a Saúde Pública. Os métodos e modelos de processo de ciência de dados, especificamente o KDD, SEMMA e CRISP-DM, foram explorados criticamente. Avaliando seus pontos em comum, um novo modelo de processos denominado KDD-PH (*Knowledge Discovery in Databases for Public Health*) foi proposto, sugerindo etapas específicas para um modelo de processos de *ciência de dados* para a pesquisa em saúde pública. Produções acadêmicas autorais foram apresentadas como resultados da aplicação prática destes métodos.

Palavras-chave: Ciência de Dados. Métodos epidemiológicos. Saúde Pública.

Abstract

Terms such as big data and data science have gained visibility in recent decades, being generally associated with the challenges of analyzing large databases. The aim of this work is to gather theories and methods of Health Data Science as a contribution to Public Health, studying the cycle of generation and dissemination of health information and apply techniques of data collection, extraction and visualization with science methods. Starting from a historical perspective, the relationship between data and health was visited, presenting a new paradigm of data science and health, considering the hybrid possibilities of a theory & data driven science for Public Health. Data science process methods and models, specifically KDD, SEMMA and CRISP-DM, have been critically explored. Evaluating their common points, a new process model called KDD-PH (Knowledge Discovery in Databases for Public Health) was proposed, suggesting specific steps for a data science process model for public health research. Authorial academic productions were presented as a result of the practical application of these methods.

Keywords: Data Science. Epidemiological methods. Public health.

Lista de ilustrações

Figura 1 – Selo da *Statistical Society on London (1838 – 1857)* 16
Figura 2 – O processo KDD 42
Figura 3 – Objetivos de *data mining* 49
Figura 4 – O método KDD como um processo iterativo 53
Figura 5 – Modelo de processos SEMMA revisto (2016) 55
Figura 6 – Modelo de processos CRISP-DM 56
Figura 7 – Modelo de processos KDD-PH 59

Lista de tabelas

Tabela 1 – Volume de alguns sistemas de informação de saúde nacionais 44
Tabela 2 – Número de municípios nos censos demográficos, segundo as grandes regiões brasileiras 47

Lista de quadros

Quadro 1 – Eras da evolução da epidemiologia 32
Quadro 2 – Os quatro paradigmas da ciência 38
Quadro 3 – Correspondências entre os modelos de processo KDD, SEMMA e CRISP-DM 58

Lista de abreviaturas e siglas

ANA	Agência Nacional das Águas
ANM	Agência Nacional de Mineração
API	*Application Programming Interface*
APAC	Autorizações de Procedimentos de Alta Complexidade
AIH	Autorizações de Internações Hospitalares (AIH)
CDTS	Centro de Desenvolvimento Tecnológico em Saúde
COVID-19	*Corona Virus Disease 2019*
CIDACS	Centro de Integração de Dados e Conhecimento para Saúde
CID-10	Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde, 10a edição
CRISP-DM	*Cross-Industry Standard Process for Data Mining*
DataSUS	Departamento de Informática do SUS
DEXL Lab	*Data Extreme Lab*
DNPM	Departamento Nacional de Produção Mineral
DVH	Doenças de Veiculação Hídrica
ETL	*Extraction, Transform and Load*
FTP	*File Transfer Protocol*
IBGE	Instituto Brasileiro de Geografia e Estatística
IHME	*Institute of Health Metrics and Evaluation*
INPE	Instituto Nacional de Pesquisas Espaciais
IRD	*Institut de recherche pour le développement*
KDD	*Knowledge Discovery in Databases*
KDD-PH	*Knowledge Discovery in Databases - Public Health*
LNCC	Laboratório Nacional de Computação Científica

MAUP	*Modifiable Areal Unit Problem*
MS	Ministério da Saúde
ODKD	*Ontology Driven Knowledge Discovery*
PCDaS	Plataforma de Ciência de Dados aplicada à Saúde
PNAD	Pesquisa Nacional de Amostra de Domicílios
PNS	Pesquisa Nacional de Saúde
POF	Pesquisa de Orçamentos Familiares
RESP	Registro de Eventos em Saúde Pública
SciELO	*Scientific Electronic Library*
SEMMA	*Sample, Explore, Modify, Model, Assess*
SIA	Sistema de Informações Ambulatoriais
SIAB	Sistema de Informações de Atenção Básica
SIH	Sistema de Informações Hospitalares
SIM	Sistema Nacional de Mortalidade
SINASC	Sistema de Informações de Nascidos Vivos
SINAN	Sistema Nacional de Agravos de Notificação
SISMAMA	Sistema de Informação do câncer de mama
SISCOLO	Sistema de Informação do câncer do colo do útero
SISAGUA	Sistema de Informação de Vigilância da Qualidade da Água para Consumo Humano
SIVEP	Sistemas de Vigilância Epidemiológica
SQL	*Structured Query Language*
NoSQL	*Not Only Structured Query Language*
SUS	Sistema Único de Saúde

Sumário

1 INTRODUÇÃO 14
1.1 Objetivos 18
1.2 Estrutura da tese 18

2 REFERENCIAL TEÓRICO 19
2.1 Informação quantitativa em saúde: uma perspectiva histórica 19
2.2 Ciência de dados em saúde: um novo paradigma 33

3 METODOLOGIA 40
3.1 KDD 40
3.2 SEMMA 54
3.3 CRISP-DM 56
3.4 Pontos em comum entre os modelos de processos 57
3.5 KDD-PH: um modelo de processos para mineração de dados em Saúde Pública 58

4 RESULTADOS 63
4.1 Construção teórica 64
4.2 Captura e seleção de dados 80
4.3 Do constructo teórico à mineração 94
4.4 Da coleta distribuída à visualização 109
4.5 Da ideia, os desafios metodológicos e tecnológicos à visualização de dados 125

5 CONCLUSÃO 154

REFERÊNCIAS 157

1 Introdução

Novos modos de pensar emergem periodicamente e desafiam as teorias e abordagens previamente aceitas pela comunidade científica, ao passo que a ciência dominante não consegue mais responder a todas as perguntas e requer a formulação de novas ideias (KUHN, 1970). A história da análise de dados reflete este comportamento, passando por diversos movimentos e mudanças de paradigmas. Este mesmo comportamento de mudança e evolução pode ser observado especificamente na análise de dados em saúde (ALMEIDA-FILHO, 1986; KITCHIN, 2014).

Termos como *big data* e *ciência de dados* ganharam visibilidade nas últimas décadas, sendo em geral associados aos desafios de análise de grandes bases de dados. Relatórios, reportagens para o grande público, editoriais e artigos científicos foram produzidos sobre este tema, procurando conceituar estes termos e avaliar os possíveis impactos na sociedade e academia (KHOURY; IOANNIDIS, 2014).

De modo geral, as definições de *ciência de dados* e *big data* convergem em algumas propriedades. Pode-se destacar que *ciência de dados* é um campo interdisciplinar para o estudo de dados estruturados ou não estruturados, que podem ser pequenos ou grandes em volume e terem outras propriedades, como grande velocidade de produção e variedade de conteúdo, podendo ainda serem estáticos ou alimentados em tempo real. As atividades de *ciência de dados* incluem a extração de dados, sua preparação, exploração, transformação, armazenamento, recuperação, manutenção das infraestruturas computacionais necessárias, aplicação de diversas estratégias de mineração de informação e aplicação de métodos de aprendizado de máquina, tal como o emprego de recursos de apresentação dos resultados e predições (HAYASHI, 1998; PRESS, 2013; KHOURY; IOANNIDIS, 2014; AALST, 2016; HOUSEH; KUSHNIRUK; BORYCKI, 2019).

Pode-se refletir que o campo de *ciência de dados* apresenta, como ciência, certa sobreposição com a *estatística*. Como Aalst (2016) afirma, o campo de *ciência de dados* se origina da *estatística*, que também deu origem ou impulsionou vários outros campos. Esta aparente sobreposição de campos levantou a discussão sobre *ciência de dados* ser um novo campo científico ou uma evolução da *estatística*. Donoho (2017) apresenta lúcida discussão sobre o tema.

Huber (2011) argumenta que a análise de dados se ocupa em analisar dados de todos os tipos, através de qualquer método que se apresente viável. Ao se considerar a *estatística* como a arte de coletar e interpretar dados, desde o planejamento da coleta até a apresentação dos resultados, pode-se afirmar que a *ciência de dados* faz parte da *estatística*.

Contudo, os métodos da *ciência de dados* não são puramente *estatísticos* pelo ponto de vista mais tradicional, com origens fundamentadas na matemática e probabilidade *a priori*. Vários métodos de *ciência de dados* são inferenciais e dedutivos, partindo da amostra para a população, mas a *ciência de dados* também utiliza métodos indutivos, enquanto que a *estatística*, do ponto de vista mais tradicional, defini-se pela inferência da parte para o todo.

Tukey já previa a necessidade de uma profunda ampliação da *estatística* e o surgimento do campo de "análise de dados" (TUKEY, 1962; TUKEY, 1977).

> *All in all, I have come to feel that my central interest is in data analysis, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data. (TUKEY, 1962)*

Esta diferenciação entre *estatística* e análise de dados – que posteriormente passa a se chamar *ciência de dados*– se explica parcialmente por sua preocupação com os objetivos da *estatística*. Para Tukey, a *estatística* passava a se preocupar demasiadamente com a precisão matemática, abstrata e teórica, se afastando das necessidades mais próximas de métodos que compreendessem todo o processo de análise de dados e as demandas da sociedade por informação.

A primeira versão do selo da *Statistical Society of London*, fundada em 1838 e berço da ciência estatística, representa de certo modo este conceito. O selo se constitui de um fecho de trigo com o lema "*aliis exterendum*" ("para ser interpretado por outros"). Em sua origem, a *estatística* tem por objetivo reunir fatos de modo imparcial, para que outras ciências os interpretassem, se afastando da influência da Política e do Estado (HILTS, 1978).

Conforme argumenta Shmueli (2010), a *estatística* apresenta uma abordagem tradicionalmente explanatória, onde busca-se testar uma teoria causal: dado um modelo teórico causal, modelos estatísticos são aplicados nos dados de modo a se testar a hipótese de causalidade. Desta forma, as hipóteses de pesquisa são dadas em termos de constructos teóricos.

> *Only after the theoretical component is completed, and measurements are justified and defined, do researchers proceed to the next step where data and statistical modeling are introduced alongside the statistical hypotheses, which are operationalized from the research hypotheses (SHMUELI, 2010).*

Ao observar a utilização a evolução da *epidemiologia* e atual aplicação dos conceitos de *big data* e *ciência de dados* em *Saúde Pública*, pode-se notar certa semelhança no embate de campos e conceitos. Artigos e editoriais foram lançados apontando para as

Figura 1 – Selo da *Statistical Society on London (1838 – 1857)*

Fonte: extraído de Hilts (1978)

possibilidades de ganho da *Saúde Pública* em incorporar recursos de *big data* aos métodos epidemiológicos tradicionais (HERLAND; KHOSHGOFTAAR; WALD, 2014; FUNG; TSE; FU, 2015; MOONEY; WESTREICH; EL-SAYED, 2015; MOONEY; PEJAVER, 2018; VAYENA et al., 2018; DICLEMENTE et al., 2019; HOUSEH; KUSHNIRUK; BORYCKI, 2019; CHIAVEGATTO-FILHO, 2015).

Como argumenta Ford et al. (2019), com o crescimento da complexidade dos problemas de *Saúde Pública*, se faz necessária uma contínua revisão do campo, tornando-o capaz de aproveitar a oferta de dados e a diversidade de métodos disponíveis. Novas fontes de dados passam a requerer novos métodos de estudo e abordagens diferenciadas, enfrentadas por equipes transdisciplinares (DICLEMENTE et al., 2019).

Denominando esta tendência como *epidemiologia digital*, Boilson et al. (2018) e Hay et al. (2013) argumentam que os objetivos da *epidemiologia tradicional* são mantidos, mesclando-se agora com fontes de dados eletrônicas que emergem com o advento da tecnologia da informação. (SALATHÉ et al., 2012) aponta que a disponibilidade de novos dados, em conjunto com dados tradicionais, apresenta-se como uma oportunidade única para novos sistemas de vigilância epidemiológica, que podem operar sem fronteiras internacionais, preenchendo hiatos da infraestrutura de *Saúde Pública* e complementando os métodos tradicionais de vigilância.

Dentre as diversas possibilidades metodológicas da *ciência de dados*, podem-se destacar as possibilidades de modelagem para predição. Os grandes conjuntos de dados disponíveis atualmente contêm relações e padrões muito complexos para se formular teorias e hipóteses *a priori*, mas que podem ser plenamente explorados pelos métodos de mineração de dados (SHMUELI, 2010).

Para DiClemente et al. (2019), se faz necessário abrir o campo da epidemiologia

para novas metodologias, postulados teóricos, estratégias e abordagens que representem um pensamento transdisciplinar. Ao se considerar o caráter multifacetado da definição de saúde proposta pelo campo de *Saúde Coletiva* (PAIM; De Almeida Filho, 1998; SCLIAR, 2007), a *Ciência de Dados em Saúde* se revela como um campo primariamente transdisciplinar. Desta forma, o *cientista de dados em saúde* não preserva um viés disciplinar de sua formação acadêmica, mas passa a trabalhar em grupos transdisciplinares que reúnem habilidades investigativas, clínicas, de governança e gerenciamento de recursos tecnológicos Ford et al. (2019).

Pretende-se com este trabalho reunir teorias e métodos da *Ciência de Dados em Saúde* como uma contribuição à *Saúde Pública*. Nesta tese, apresenta-se um Referencial teórico que busca traçar a evolução histórica da informação quantitativa em saúde, seguida da apresentação da Metodologia. A seguir, Resultados práticos da aplicação destes métodos no campo da *Saúde Pública* são apresentados através de artigos científicos revisados por pares. Aplicando um ou mais passos desta metodologia, busca-se assim apresentar contribuições relevantes da *Ciência de Dados em Saúde* para a *Saúde Pública*, cumprindo-se os Objetivos a seguir elencados.

1.1 Objetivos

Objetivo geral

Estudar o ciclo de geração e disseminação de informação em saúde e aplicar técnicas de coleta, extração e visualização de dados estruturados e não estruturados usando métodos de *ciência de dados*.

Objetivos específicos

1. Estudar a evolução do tratamento e visualização de dados em saúde, incluindo as etapas pré-computacionais e informacionais, até o advento da *ciência de dados*.
2. Descrever os processos modernos de tratamento e visualização de dados em alguns projetos estratégicos.
3. Aplicar modelos de processos de análise de dados em pesquisas na área da saúde.
4. Analisar criticamente as vantagens e limites do uso das metodologias de *ciência de dados*.

1.2 Estrutura da tese

Esta tese consiste na apresentação de um Referencial teórico (p. 19) e uma Metodologia (p. 40) comum de análise de dados em saúde. Em seguida, como Resultados (p. 63), são apresentados cinco artigos, dos quais três foram publicados, um foi aprovado para publicação e um será submetido, além de um capítulo de livro em fase de revisão. Por fim, uma Conclusão (p. 154) encerra o trabalho com uma revisão crítica e propositiva dos métodos utilizados.

2 Referencial teórico

> *Va, pensiero, sull'ali dorate; Va, ti posa sui clivi, sui colli, Ove olezzano tepide e molli L'aure dolci del suolo natal!*
>
> —Verdi, *Nabucco*.

Neste capítulo pretende-se apresentar as principais conceituações teóricas necessárias para a construção deste trabalho. A primeira seção trata de uma breve perspectiva histórica da formação da informação quantitativa em saúde, de seus primórdios até o paradigma predominante de meados do século XX. A segunda seção procura refletir teoricamente sobre o estado atual da informação quantitativa em saúde com base no paradigma do *big data*.

2.1 Informação quantitativa em saúde: uma perspectiva histórica

Pretende-se com este capítulo apresentar uma breve revisão das origens dos usos e aplicações da informação em saúde, com um recorte dos métodos de coleta e análise de dados quantitativos em uma perspectiva histórica. Considera-se aqui a evolução deste campo no Ocidente.

Para esta revisão histórica, as obras de Rosen (1993), Susser e Susser (1996a), Susser e Susser (1996b), Susser e Stein (2009) se mostraram importantes fontes de conhecimento e organização. Contudo, cabe ressaltar que estes movimentos, aqui agrupados linearmente e em sequência, ocorreram no decorrer de diversas décadas, com avanços e retrocessos, frutos de diversas iniciativas até mesmo discordantes ou sobrepostas, ora objetivando certas doenças e epidemias, ora a Saúde Pública como um todo.

Da história antiga a Galeno

Conforme afirma Rosen (1993), os principais problemas enfrentados pelo homem se relacionam à vida comunitária, como no controle de doenças transmissíveis, melhorias das condições sanitárias e alimentação. Em geral, estes problemas são relacionados a uma dificuldade dualística de separar o saudável do insalubre, como separar a água limpa da não potável ou o alimento comestível do estragado, por exemplo.

De semelhante modo, as doenças transmissíveis também sempre remeteram a uma lógica de separação, onde se procura o afastamento do indivíduo doente do restante da sociedade via o isolamento físico. Para que a comunidade consiga se proteger efetivamente, torna-se necessário que os casos de certas doenças sejam reportados às autoridades locais.

> *When people suffering from transmissible diseases may directly menace the health of those around them, the community acting through its institutions feels justified in subjecting the individual to restraints and even sanctions in order to protect itself. Thus, people suffering from certain communicable diseases have had to be reported to the authorities (ROSEN, 1993, p. 26).*

Este ato de identificar quais indivíduos estão doentes e reportá-los para autoridades ou instituições pode ser visto como um embrião da vigilância sanitária em saúde. Ao se apontar quais pessoas estão doentes e destas, quais devem ser separadas do convívio da sociedade, cria-se uma necessidade básica de identificar, enumerar e caracterizar estes indivíduos.

Exemplos mais diretos desta lógica de separação e comunicação dos casos a autoridades locais podem ser retirado das passagens bíblicas, onde relatam-se os casos de hanseníase e deficiências físicas, e os mecanismos sociais de lidar com estes casos. De semelhante modo, casos hanseníase na Idade Média passavam pelo crivo de uma comissão formada por um bispo, outros clérigos e um *leproso* tido como "especialista" da doença. Após ser diagnosticado positivamente, o caso era comunicado as autoridades locais e sanções eram impostas ao doente, incluindo o seu isolamento para fora dos muros da cidade.

Também na Idade Média, os Estados em formação já tomavam para si a responsabilidade de executar medidas para o controle de doenças contagiosas como a peste, criando cordões sanitários entre comunidades e instituindo postos de observação, hospitais de isolamento, procedimentos de desinfecção e outras medidas de quarentena para casos suspeitos. A sociedade medieval já apresentava um "maquinário administrativo" para prevenção de doenças e proteção da saúde na comunidade(ROSEN, 1993).

Em uma sociedade ainda dominada por práticas supersticiosas e de conhecimento científico esparso, a existência e persistência de doenças era fortemente ligada à noção de pecado e punição. Deste modo, é natural que a Igreja assumisse o papel de autoridade local mais forte para conduzir assuntos relacionados a doença e morte.

Para além da noção religiosa das doenças, a explicação científica sobre elas se concentrava basicamente nas teorias hipocráticas. Galeno, proeminente médico romano, sistematizou a "teoria miasmática", onde doenças eram causadas por alterações atmosféricas, pela corrupção do ar causada pela decomposição da matéria orgânica e por águas pútridas. Esta doutrina persistiu na medicina europeia por cerca de 1.500 anos, sendo também aceita pela Igreja (SUSSER; STEIN, 2009; ROSEN, 1993).

Naturalmente, não havia no período medieval um registro quantitativo formal dos casos de doenças, muito menos algo que servisse para o monitoramento contínuo da saúde da população como um todo. Os casos esporádicos de doenças transmissíveis eram simplesmente manejados pela própria população por seu sistema de crenças, enquanto que,

se a quantidade de casos de avolumasse ao ponto de chamar a atenção dos governos, as autoridades tomavam medidas mais drásticas como quarentenas, fechamento de portos e barreiras comerciais.

Aritmética política e a era das estatísticas sanitárias

A mudança deste cenário se inicia apenas no século XVI, com o Renascimento e o Mercantilismo. Com o fortalecimento dos estados-nação e busca por riquezas através das grandes expedições, os setores de comércio e industria se fortaleceram, levando a formação de cidades maiores, fora da estrutura feudal.

Neste momento, o Estado passa a se preocupar não apenas com suas forças militares, mas também diretamente com sua força produtiva. Mais especificamente, o Estado passa a coletar e registrar seus números e a medir a sua capacidade de produzir bens e riquezas. A maquina pública toma corpo e, no intuito de manter esta capacidade de produção, o Estado assume a responsabilidade de administrar questões relacionadas a saúde da comunidade.

Grandes avanços neste sentido são liderados pela Inglaterra, mas também seguidos pela França e Alemanha em diferentes graus. Na Inglaterra do século XVII surge o conceito da "aritmética política" com os trabalhos de William Petty (1623 – 1687), apresentando direcionamentos para o Estado reunir dados básicos da sociedade. Em Dublin, Petty coletou e analisou dados sobre óbitos e suas localizações, tentando-os relacionar com fatores ambientais. Conforme aponta Susser e Stein (2009), "suas ideias foram significantes e influenciaram especialmente economistas nas áreas de saúde e seguros", por ter criado também métodos para calcular o custo da perda prematura de vidas.

Contudo, ainda que Petty já reconhecesse a importância dos dados de saúde, é com John Graunt (1620 – 1674) que as primeiras contribuições sólidas para a saúde pública foram realizadas, através de seu meticuloso trabalho de coleta de causas de óbito em Londres.

Graunt, com a ajuda de uma equipe formada por mulheres, reúne dados sobre a causa de morte aparente da população de Londres a partir dos registros paroquiais, organizando-as em um compilado estatístico de dados semanais sobre mortalidade e suas principais causas, chamado "*Natural and Political Observations made upon the Bills of Mortality*".

Além disso, Graunt idealizou um sistema de alerta para epidemias de peste bubônica, que não chegou a ser implantado. Ainda assim, suas *tábuas de mortalidade* contribuíram para as primeiras estimativas estatísticas da população de Londres e são um recurso histórico ímpar de detalhes da demografia da população londrina.

Dentre outras descobertas, Graunt identificou que certas doenças apresentavam uma proporção relativamente constante no total de óbitos no decorrer do tempo, reconheceu

também que a proporção de óbitos masculino tende a ser maior que a do sexo feminino e que a cidade apresenta um número proporcionalmente maior de óbitos do que o campo.

A significância dos trabalhos de Graunt se dá pelo emprego direto de métodos estatísticos, ainda que rudimentares, para análise de dados de saúde e por ter classificado os óbitos em grupos. Pode-se também afirmar que seu trabalho foi pioneiro no que se chama hoje de "análise de dados secundários", por ter utilizado dados coletados por outras pessoas (no caso, paróquias) (SUSSER; STEIN, 2009).

Sobre a qualidade dos dados, Graunt aponta que estes devem ser interpretados seguindo suas limitações de coleta e precisão, levando-se em conta, por exemplo, oscilações que podem ser causadas em momentos de epidemia. Graunt destaca que, no intuito de evitar propagar a notícia de uma epidemia de praga, as paróquias tendiam a classificar erroneamente estes óbitos. Ainda mais, Graunt sugere que dados defeituosos ou enviesados ainda podem ser interpretados de uma maneira útil para a sociedade, desde que se empregue um método lógico e honesto que abrace as características dos dados.

Interessante notar que os livros de registro de casamentos, nascimentos e óbitos das paróquias religiosas não foram criados para a finalidade estatística de mensurar a população, mas para a formalização dos assuntos religiosos de seus fiéis. Desta forma, os estudos quantitativos em saúde surgem a partir de dados secundários, coletados para outras finalidades.

Com Petty e Graunt durante o século XVIII, avanços para a quantificação específica de doenças foram esporádicos. Ainda assim, o Iluminismo e a Revolução Francesa criou terreno para autores de métodos estatísticos como Pascal, Bernoulli, Fermat e Huygens na área de probabilidade e incerteza, e para Leibniz no cálculo. Ainda que estes não tivessem se dedicado diretamente aos dados de saúde, já endereçavam sobre a possível importância do desenvolvimento destes métodos para a saúde. (ROSEN, 1993; SUSSER; STEIN, 2009)

Conforme aponta Trunk (2006), Leibniz recomenda que todos os dados referentes a tentativas de cura e de mortalidade deviam ser coletados, tanto nas grandes cidades como nas áreas rurais. De fato, Leibniz destaca a importância da medicina popular:

> *This data raising starts with the collection of all popular medical knowledge, especially of the healing secrets of old women. Anyone who gives details about a credible way of medical treatment shall be honoured. The village physician who collects a great amount of interesting knowledge will be decorated with honours as well. Anyway, physicians are obliged to record everything of importance. (TRUNK, 2006, p. 3)*

Conforme aponta Rosen (1993), a "aritmética política" de Petty é um meio para um fim, especificamente, o meio para se obter poder e prosperidade nacional. Desta forma, a evolução natural deste princípio é ir além da quantificação da saúde da nação para uma tentativa de real atuação sob as condições de vida que impossibilitavam a promoção

da saúde, prevenção de doenças e acesso ao cuidado. Este avanço implica no criação de uma "política de saúde", com objetivos, método e parâmetros observáveis para o acompanhamento de sua implantação.

Com William Petty, a Inglaterra surge como liderança em avanços no campo da saúde pública. Petty compreende que certas questões de saúde devem ser assumidas estrategicamente pelo Estado e propõe diversas medidas, como a prevenção de doenças contagiosas em crianças como um modo de salvaguardar a população, a necessidade de fomento de pesquisas médicas pelo Estado e o acompanhamento da morbidade e mortalidade em certas doenças ocupacionais. Destaca-se a sugestão de que as análises das necessidades na área da saúde fossem realizadas pelo Estado, sendo conduzidas seguindo métodos estatísticos – como os usados por John Graunt – para o cálculo do número de médicos, leitos e cirurgiões necessários.

Contudo, apesar das grandes potencialidades das ideias surgidas nesta época, poucas foram executadas devido a tendências políticas e administrativas contrárias. Pode-se afirmar que na Inglaterra do século XVIII, os problemas de saúde ainda eram largamente administrados em uma escala local e provinciana. Apenas no século XIX, com o advento da era industrial, é que as soluções para a organização dos problemas de saúde tornaram-se uma preocupação nacional e as propostas de William Petty e outros foram recuperadas e implementadas em alguma medida.

Com o crescimento das grandes cidades industriais no século XIX, evidências de suas consequências sob a saúde começam a aparecer na Inglaterra na forma estatística através do Censo Decenal (iniciado em 1801) e no registro compulsório de óbitos, nascimentos, casamentos e óbitos (iniciado em 1837). Estes primeiros dados já apontavam para problemas das condições de vida da sociedade industrial.

> *Filth, disease, destitution, and the demand for a reduction in the burden of poor relief are the roots from which the Sanitary reform sprang (ROSEN, 1993, p. 116).*

Na Inglaterra, o avanço da aritmética política e seus reflexos na saúde passa neste tempo, obrigatoriamente, pelo nome de Edwin Chadwick. Advindo da escola de "filósofos radicais" de Bentham (ROSEN, 1993), ele propunha a aplicação de métodos científicos na administração do Estado. Chadwick atuou na edição do "*Poor Law Amendment Act*" (1834) e o relatório produzido procurou provar a relação das doenças – principalmente das transmissíveis – com más condições de vida, ausência de abastecimento e saneamento de água e limpeza urbana. Chadwick apontou que as principais razões para estes problemas estava na má administração pública, que negligenciava o ato de legislar e fiscalizar sobre estes temas (ROSEN, 1993). De modo inovador, Chadwick se baseava nas inovações de técnicas e análises estatísticas em suas afirmações e argumentos.

O conceito de saúde pública toma uma nova conotação, deixando de ser um problema relacionado a ciência médica e a prática clínica, tornando-se um *problema de engenharia*.

Isto pode significar um aparente paradoxo para o pensamento da época: resolver um problema de saúde através de métodos não advindos da medicina. Este novo conceito evidenciou a necessidade do Estado em considerar o conhecimento científico existente para medidas preventivas, de maneira eficiente e consistente, para a eliminação dos problemas de saúde que prejudicavam a sua capacidade produtiva.

Segundo o mesmo relatório:

> *The great preventives, drainage, street and house cleasing by means of supply of water and improved sewerage, and specially the introduction of cheaper and more efficient modes of removing all noxious refuse from the towns, are operations for which aid must be sought from the science of the Civil Engineer, not from the physician, who has done work when he has pointed out the disease that results from the neglect of proper administrative measures, and has alleviated the suffering of the victims. Chadwick apud Rosen (1993, p. 120).*

Neste sentido, a saúde da população é entendida como uma característica de sua coletividade. As doenças passam a serem estudadas não apenas por uma perspectiva clínica de adoecimento, sintomas, diagnóstico, tratamento e cura, ligado ao caso clínico do indivíduo isoladamente. Forma-se um ponto de vista tomado à partir do retrato de uma *coleção de indivíduos doentes*. Este processo, nomeado por Guèrin em 1838 como Medicina Social, toma coletivamente a questão da saúde (NUNES, 1998). As doenças passam a ser estudadas por um método numérico que quantifica, mede e compara, fortemente ligado a ciência Estatística, também em ascensão.

Os argumentos de Chadwick levaram a implantação de legislações que sustentavam políticas sociais de renda mínima, de saneamento, habitação e proteção do trabalhador na Inglaterra, através da melhoria relativa das condições de saúde da população. O principal legado destas leis foi a formação de um mercado de trabalho adequado para o funcionamento do sistema econômico, através de medidas que atuavam sobre a saúde da população e na reconfiguração da máquina pública através de uma maior centralização, uniformidade e eficiência.

Essa etapa histórica marca também a adoção de uma visão pragmática da saúde, isto é, mesmo desconhecendo as causas das doenças, era possível identificar os grupos populacionais mais afetados e promover ações que minimizassem o sofrimento, pobreza e a carga de doenças sobre os mais pobres.

De fato, estas novas legislações sobre saúde e saneamento foram oriundas de uma variedade de forças econômicas e sociais, que perceberam que as doenças endêmicas e

epidêmicas eram um problema para toda a comunidade, levando a uma preocupação crescente sobre os custos envolvidos na incapacitação e perda de mão de obra. Conforme aponta John Simon em 1858:

> *Sanitary neglect is mistaken parsimony. Fever and cholera are costly items to count against the cheapness of filthy residence and ditch-drawn drinking-water: widowhood and orphanage make it expensive to sanction unventilated work-places and needlessly fatal occupations... The physical strength of a nation is among the chief factors of national prosperity. Apud Rosen (1993, p. 127).*

Tais objetivos na máquina administrativa não seriam possíveis sem o acompanhamento estatístico das condições de saúde da população. Deve-se mencionar neste período as contribuições de William Farr (1807 – 1883), com seus relatórios estatísticos sobre doenças nas casas, nas fábricas e nas comunidades, servindo de embasamento empírico para a proposição destas novas leis. Farr encabeça a criação de um sistema nacional de estatísticas, que posteriormente serviu de base para os trabalhos de Edwin Chadwick, Friedrich Engels, Florence Nightingale, John Simon e outros.

A evolução dos métodos e aplicações da informação quantitativa em saúde, e em especial da Epidemiologia, é vinculada diretamente com as necessidades da saúde da sociedade. Desta forma, sua formação não nasce de uma idealização acadêmica de campo de pesquisa, com uma forma, ordem e regras pré-estabelecidas, mas se guia – e se limita – pelas razões econômicas, sociais e políticas dos tempos.

De fato, a persistência e sucessão de conhecimentos acompanha a própria evolução da ciência médica. Teorias da fundação da ciência grega, dos romanos e das civilizações islâmicas foram sendo lentamente incorporadas, miscigenadas e, por fim, superadas por novas descobertas. A evolução destes conceitos leva a criação de um conhecimento cumulativo, que necessita de constante atualização e propagação, onde teorias que se provaram ineptas cedem em favor das novas descobertas. Contudo, esta produção e difusão do conhecimento se dá em fases lentas, com camadas de sobreposição no tempo (SUSSER; STEIN, 2009).

Com Farr, os dados de saúde deixam de ser simplesmente oriundos de uma aplicação secundária dos registros de casamento, nascimento e óbito, e passam a ser coletados diretamente com o intuito de acompanhamento da saúde da população possibilitando usos mais precisos e específicos como o de Friedrich Engels. Em 1885, Engels publica o relatório "A Situação da Classe Trabalhadora na Inglaterra", estratificando dados de mortalidade por diversos recortes, como áreas ruais e cidades industriais, bairros operários e burgueses, casas em pior estado e casas adequadas.

Ainda que considerando os avanços nas estatísticas de análise, no próprio processo de coleta de dados, e já considerando relações entre pobreza, ausência de assistência e saúde, toda a lógica da saúde pública deste período ainda se fundava na teoria miasmática.

Deste modo, as medidas preventivas se voltavam para melhorias no abastecimento de água, esgotamento e hábitos sanitários (SUSSER; STEIN, 2009).

Esta "Era das Estatísticas Sanitárias" é superada apenas com os avanços da "Teoria dos Germes". Alguns autores criaram teorias e proposições de que as doenças eram transmitidas através de algum agente, o que fundamentalmente negava as teorias miasmáticas. Estas ideias emergiram pontualmente durante séculos, concorrendo com a teoria miasmática. Em geral, não ganhavam aceitação frente a popularidade e a facilidade com a qual a teoria miasmática explicava os fenômenos da época.

A teoria dos germes e a era das doenças infecciosas

Girolamo Frascatoro (1478 – 1553), recupera a ideia já existente de que uma doença pode ser transmitida de uma pessoa para outra e idealiza que esta transmissão se dava através de um *germe* ou *semente*, em um trabalho publicado em 1546. Athanasius Kircher (1602 – 1680), com o uso de um microscópio rudimentar, identificou pequenos corpos vivos em pessoas com doenças contagiosas. Contudo, se assumia que estes germes surgiam de um processo de geração espontânea, conforme o conhecimento geral da época[1]. Kircher aponta que estes germes causavam putrefação dentro dos corpos mas morriam quando expostos ao ar. Diversos outros trabalhos surgiam esporadicamente apontando para a existência de germes e evidências de sua capacidade em transmitir doenças, como Peter Ludwig Panum (1820 – 1885), Alexander Gordon (1752 – 1799), Ignác Semmelweis (1818 – 1865), Joseph Lister (1827 – 1912), John Snow (1813 – 1858) e William Budd (1811 – 1880) (SUSSER; STEIN, 2009; SUSSER; SUSSER, 1996a).

Estas ideias e teorias da presença de um germe capaz de transmitir doenças, ainda que não aceitas amplamente, foram formando uma corrente crescente de descrença da teoria miasmática. Em resposta a isto, Thomas Sydenham tenta incorporar a existência de germes na teoria miasmática, onde nas estações da primavera e outono os miasmas apresentavam estes germes. Incrivelmente, este acréscimo era plausível, ao passo de que o número de casos de doenças contagiosas aumentava nestas estações. Contudo, algumas pesquisas levaram a conclusões mais diretas contra a teoria miasmática. Dentre elas, destaca-se abaixo o trabalho de John Snow.

Apesar de ser mais conhecido hoje pelo seu trabalho nas epidemias de cólera, o médico John Snow ganhou evidência em sua época por seus estudos no campo de anestesia, tendo inclusive realizado este procedimento na Rainha Vitória. O interesse de Snow pela cólera remete a época de quando trabalhava como médico em minas de carvão em sua cidade natal de Tyneside, onde ocorreu uma epidemia de cólera em 1832.

Novas epidemias de cólera, agora em Londres, chamaram novamente a atenção

[1] Francisco Redi (1626 – 1698) vem negar esta teoria posteriormente.

de Snow, levando-o a acompanhar três epidemias (1832, 1849 e 1854). Estas epidemias o levaram a três conclusões: (1) os padrões de distribuição da doença e seu modo de agir apontam para a ingestão de água contaminada; (2) o agente de contaminação era um organismo e (3) este organismo tem um tamanho microscópico, possivelmente inanimado, mas capaz de transmitir a doença e se replicar.

Nos estudos das duas primeiras epidemias, Snow conseguiu relacionar os casos de cólera com a presença de fossas inapropriadas e próximas a rede de abastecimento de água. O terceiro estudo (sobre a epidemia de 1854) é o mais conhecido, relacionando a contaminação à "*Broad Street Pump*", que era abastecida pela água do Rio Thames à jusante de uma área de despejo de esgoto.

Seus estudos contestavam diretamente as teorias miasmáticas da época, apontando para o contágio através da ingestão de água contaminada e não por odores e gases atmosféricos inalados. Deste modo, suas investigações sobre cólera necessitava de um grande rigor científico para fazer frente a teoria mais aceita e ao pensamento comum da época.

Na epidemia de 1854, Snow coletou em campo a localização geográfica de cada caso, data e horários referentes ao início dos sintomas, outras características do quadro clínico dos doentes, e dados sobre o consumo de água e sua fonte de abastecimento. Como nos estudos anteriores, ao seguir a rota do abastecimento de água, Snow encontrou novamente fortes indícios da relação entre água contaminada e casos de cólera.

Além de contribuir com avanços sobre o contágio de doenças, o trabalho de Snow é inovador por seu nível de detalhamento, pela localização dos casos em mapas e a utilização de princípios de controle[2]. Enquanto os estudos anteriores de cólera se davam em escalas menos detalhadas, agregando-se casos em unidades de análise maiores, Snow iniciou suas conclusões na escala individual, que se mostrou adequada a doença e ao seu modo de contágio. Desta forma, suas conclusões, iniciadas nesta escala, podiam ser estendidas a escalas menos refinadas de análise. Ainda assim, suas contribuições foram recebidas sem grande alarde por contrariar a teoria vigente miasmática. Apenas 30 anos depois, com Pasteur e Koch, o vibrião do agente da cólera foi descrito e seus trabalhos foram plenamente aceitos.

Em 1840, Jacob Henle publicou um artigo reunindo ideias e proposições sobre doenças e germes, levantando a hipótese de que a infecção através de organismos diminutos eram uma grande causa de doenças. Louis Pasteur demonstrou em 1865 que diminutos organismos vivos, visíveis apenas através de microscópios, eram os responsáveis por uma epidemia que atingia os bicho-de-seda na época. Em 1882, Robert Koch estabelece que a causa da cólera é a presença de uma bactéria.

[2] Além dos casos positivos de cólera, Snow procurava explicar a não ocorrência da doença em domicílios cercados de casos.

Com estes estudos, inaugura-se a chamada era da "Teoria dos Germes" (SUSSER; STEIN, 2009), dominante entre o século XIX e meados do século XX. Agentes específicos de doenças eram identificados através da coleta e crescimento de culturas. Experimentos eram realizados sobre o modo de transmissão destes microrganismos e sobre as lesões que estes provocavam. Com o isolamento destes agentes específicos de doenças, o próximo passo lógico era no sentido da prevenção através de vacinas e cura através do uso de agentes químicos e antibióticos específicos.

A aparente facilidade de isolamento de microrganismos e a descoberta do agente causador de uma grande quantidade de doenças levou a sociedade médica a um estágio eufórico, onde imaginava-se que a prevenção e cura de todas as doenças estava ao alcance. Aparentemente, tudo poderia ser resolvido através de um laborioso trabalho de isolamento de agentes, estabelecimento de seu modo de transmissão e criação de uma vacina.

Esta euforia acabou por prejudicar imensamente os avanços anteriormente conquistados nos campos da epidemiologia e estatísticas de saúde. Com a aparente facilidade de prevenir e curar doenças através do isolamento de seus agentes causadores, os inquéritos de campo e as estatísticas oficiais de doenças tiveram sua importância esvaziada. Os pesquisadores da tradição da epidemiologia social perderam prestígio e poder na hierarquia médica.

Pode-se mencionar que alguns trabalhos passaram a surgir a partir de 1914, apontando que certas doenças eram causadas pela deficiência nutricional e não por um microrganismo, onde pode-se citar Joseph Goldberger e Edgar Sydenstricker. Este último ainda apontou que as deficiências nutricionais acabavam por perpetuar o ciclo de pobreza dos trabalhadores do ciclo do algodão nos Estados Unidos da América (SUSSER; SUSSER, 1996a). Contudo, a concentração de investimentos na pesquisa de agentes causadores de doenças eclipsou pesquisas que avançavam em direções diferentes.

Doenças crônicas e a era da caixa preta

Se faz necessário reconhecer que a Teoria dos Germes contribuiu para uma dramática redução dos óbitos causados por doenças transmissíveis nos países desenvolvidos como, por exemplo, o desenvolvimento de vacinas e antibióticos. Contudo, após a Segunda Guerra Mundial, os registros de mortalidade apontavam para o crescimento na quantidade de óbitos por doenças que não eram determinadas por um agente causador específico, mas cujo óbito era consequência da história natural de doenças crônicas. Nesta transição epidemiológica, as maiores proporções de óbitos passam a serem causadas por úlceras estomacais, doenças coronarianas e câncer de pulmão, entre outras, atingindo principalmente a população adulta de meia-idade (SUSSER; SUSSER, 1996a).

A transição epidemiológica levou a epidemiologia a uma nova era, denominada

por (SUSSER; SUSSER, 1996a) como a da "Caixa preta". Os intricados mecanismos de contagio, adoecimento e transmissão das doenças perdem foco e a busca pelos agentes causadores das doenças cessa. Com a crescente relevância das doenças crônicas, em muito ligadas aos hábitos de vida e questões hereditárias, a epidemiologia passa a criar e enfatizar novos métodos de atuação focados nestas características, em geral difíceis ou impossíveis de serem modificadas.

A obtenção de dados nesta nova epidemiologia procura se enriquecer com as chamadas "variáveis explicativas". Antes, apenas alguns dados sobre os óbitos ou adoecimento bastavam para traçar um perfil epidemiológico e descobrir associações que permitiam a busca do "agente causador". Agora, para explicar o adoecimento e poder agir em seus mecanismos, fatores associados a transmissão, adoecimento e óbito passam a serem inclusos na coleta de dados para um melhor ajuste dos modelos estatísticos. Além das típicas informações de local de nascimento, residência, contágio e óbito, as bases de dados em saúde passam a receber informações sobre hábitos de vida, trabalho, família, preferências pessoais e gênero, dentre outras.

Ainda assim, passado a afã dos novos métodos, das bases de dados maiores e análises estatísticas complexas desta nova era, reconhece-se que a "Era das Doenças Crônicas" não foi capaz de equalizar as questões de saúde, principalmente entre países desenvolvidos e em desenvolvimento com equidade. Pode-se citar, por exemplo a AIDS/HIV: o organismo de causa, as formas de transmissão e diversos fatores individuais e populacionais da epidemiologia da doença são conhecidos, contudo o "paradigma da caixa-preta" pouco conseguiu contribuir, ou muito lentamente, para o controle da epidemia.

Conforme aponta Susser e Susser (1996a)[p. 671],

> *Analysis of mass data at the individual level of organization alone, as implied by the black box paradigm, does not allow us to weigh at which points in the hierarchy of levels intervention os likely to be successfull [...] We know which social behaviors need to change, but we know little about how to change them [...] The immediate causes and the risk factors [are] known, but this knowledge [can] not be translated into protection of the public health. [...] The black box paradigm alone does not elucidate societal forces or their relation to health".*

A pesquisa em saúde deste paradigma, em sua forma mais pura, se concentra basicamente em razões de chance ou indicadores de risco, se afastando do conhecimento biológico existente. Por analisar as doenças basicamente apenas no nível individual, este paradigma acaba por não utilizar plenamente o potencial oferecido pelos novos sistemas de informações que expõem fatores de exposição, desfecho e risco em um contexto social.

Eco-epidemiologia e as caixas chinesas

Susser e Susser (1996b) delimita então uma então nova era na história da epidemiologia, chamada "Eco-epidemiologia". Segundo os autores, esta era apresenta uma complementação do "universalismo" das ciências da natureza com a "ecologia" das ciências biológicas.

Este novo paradigma reconhece que o foco nos fatores de risco no nível individual não são capazes de dar conta da complexidade das doenças, principalmente em doenças então emergentes como a AIDS. Se faz agora necessário se preocupar tanto com os caminhos causais no nível social quanto na patogênese e causalidade do nível molecular, em um sistema interativo.

Schwartz, Susser e Susser (1999) afirmam que as causas ou determinantes de saúde estão além das características individuais, que foram priorizadas pelo *paradigma do fator de risco*, devendo serem abordados segundo um sistema complexo e hierárquico.

Este sistema pode ser ilustrado por um conjunto de caixas chinesas. Estas caixas são de tamanhos diferentes e ordenados, onde uma contêm a outra. De semelhante modo, cada nível de complexidade e escala das estruturas biológicas se encaixa no outro, de nível superior, indo do nível molecular ao social.

> *Within each level, a relatively bounded structure such as nation or society or community may be characterized by lawful relations that are localized to that structure and can be discovered. At any given level within the hierarchy of scale and complexity, these lawful relations are generalizable, but only to the extent that they hold for other similar structures, whether they are societies, cities, local communities, or individuals (SUSSER; SUSSER, 1996b)[p. 675].*

Nesta estrutura multinível a epidemiologia se desenvolve em todas as escalas, da molecular a global, de modo multinível e interativo, levando em conta simultaneamente os diferentes processos de cada nível.

O quadro 1 apresenta as principais características das eras apresentadas neste capítulo e uma contribuição original: estendendo a obra de Susser, apresenta-se a seguir um novo desafio para a epidemiologia. A contínua evolução das ciências da informação levou a todos os campos do conhecimento novos desafios, essencialmente ligados a quantidade de dados disponíveis e sobre novas técnicas e metodologias que buscam, em geral, não simplificar a realidade através de modelos, mas dar conta simultaneamente de tudo o que se observa, através de dados estruturados ou não para este fim.

De fato, Susser e Susser (1996a) já antecipava as contribuições que as grandes base de dados podem dar para a pesquisa em saúde:

> *Stores of data can be mined to describe distribuitions across societies [...] Continuous accumulation of data over time can serve for overall surveillance of health states, the detection of nascent epidemics and new diseases, the response to disasters and the evaluation of interventions (SUSSER; SUSSER, 1996a)[p. 672].*

No capítulo a seguir, propõe-se um novo paradigma: *ciência de dados em saúde* e a caixa de pandora.

Quadro 1 – Eras da evolução da epidemiologia

Era	Paradigma	Procedimento analítico	Procedimento preventivo	Dados
Estatística Sanitária (primeira metade do século XIX)	Miasma: envenenamento por emanações do solo, ar e água.	Demonstrar agrupamentos de morbidade e mortalidade.	Saneamento, abastecimento canalizado, medidas sanitárias.	Dados secundários, registros de mortalidade das paróquias.
Doenças Infecciosas (final do século XIX até meados do século XX)	Teoria dos germes: agentes únicos relacionados a doenças específicas.	Isolamento laboratorial e criação de culturas, experimentos com transmissão de doenças e reprodução de lesões.	Interromper a transmissão, vacinas, isolamento dos doentes em quarentenas e hospitais específicos.	Dados primários e secundários, inquéritos e levantamentos por doenças específicas.
Epidemiologia de doenças crônicas (meados do século XX)	Caixa preta: exposição relacionada ao desfecho, sem necessidade de fatores intervenientes ou patogenia.	Razão de risco de fatores em relação a um desfecho no nível individual.	Controlar os fatores de risco modificando o estilo de vida (dieta, exercícios, etc.), agente (armas, trânsito, etc.) e meio ambiente (poluição, fumo passivo, etc).	Ênfase nos dados secundários, cruzamentos de fontes de dados sociais.
Eco-Epidemiologia (emergente)	Caixas chinesas: relações entre e através de estruturas organizadas hierarquicamente.	Análise de determinantes e desfechos em diferentes níveis de organização: dentro e entre contextos (sistemas de informação) e em profundidade (técnicas da biomedicina).	Aplicar tecnologias biomédicas e da informação, do nível contextual ao molecular.	Dados primários e secundários, dados individuais e agregados.
Ciência de dados	*big data*: métodos de predição não explicativos e não autoexplicativos.	ETL, mineração de dados, visualização de dados complexos e predição de desfechos.	Medicina de precisão, procedimentos individualizados, predição de cenários.	Grandes bases de dados estruturadas e não estruturadas, prontuários eletrônicos, internet das coisas.

Fonte: adaptado de Susser e Susser (1996a), Susser e Susser (1996b), Susser e Stein (2009), Chiavegatto-Filho (2015).

2.2 Ciência de dados em saúde: um novo paradigma

> *When they propose to establish the universal from the particulars by means of induction, they will effect this by a review of either all or some of the particulars. But if they review some, the induction will be insecure, since some of the particulars omitted in the induction may contravene the universal; while if they are to review all, they will be toiling at the impossible, since the particulars are infinite and indefinite.*
>
> —Sextus Empiricus (c. 160 CE – c. 210 CE) *Outlines of Pyrrhonism*

A produção e interpretação de dados em diversos países em desenvolvimento enfrenta dificuldades acerca da qualidade, periodicidade e granularidade necessárias para melhor orientar o processo de tomada de decisão em políticas de saúde, sendo ainda necessário o fortalecimento e avanço nas análises dos dados já existentes e coletados nestes países (BOERMA; STANSFIELD, 2007).

A produção de dados por diversos sistemas de informação e a necessidade de integração destes sistemas, tão como a inclusão de fontes de informações espaciais, demográficas e também relativas à redes sociais *on-line* trazem à tona a necessidade de aplicação de metodologias diferenciadas, apropriadas para grandes massas de dados. Neste sentido, a área de Informática em Saúde passa a assimilar os desafios das áreas de *ciência de dados* e *big data*, buscando novos referenciais para lidar com este tipo de dados (COAKLEY et al., 2013).

Tradicionalmente, as técnicas de análise de dados foram criadas para extrair informação a partir de poucos dados, em geral estáticos, limpos e de natureza pouco relacional, amostrados cientificamente e seguindo suposições claras, como independência, estacionariedade e normalidade. São dados gerados e analisados para responder uma pergunta específica, formulada anteriormente à sua produção (MILLER, 2010).

O desafio de analisar *big data* é lidar com a abundância, exaustividade e variedade, sua dinâmica, desordem e incerteza, e a necessidade de lidar com dados que não foram gerados para responder uma questão específica (KITCHIN, 2014).

O termo *big data* começou a ser utilizado no final da década de 1990 e seu significado foi se modificando conforme as demandas e a tecnologia envolvida avançava (WANG; HAJLI, 2016). Desta forma, a própria definição do termo *big data* ainda é vaga, existindo mais de 43 definições (HUANG et al., 2015). Na área de Informática em Saúde, a quantidade de dados talvez não seja a característica mais importante para a decisão quanto ao emprego do termo *big data*, ainda mais quando comparada à quantidade de dados produzida em outras áreas como financeira e de redes sociais (HERLAND; KHOSHGOFTAAR; WALD, 2014). Neste sentido, a definição proposta por Demchenko et al. (2013) parece melhor compreender os diversos aspectos da aplicação de *big data* em saúde, apresentando cinco

conceitos chaves que se iniciam com a letra "V": volume, velocidade, variedade, veracidade e valor. Pode-se ainda acrescentar a propriedade de exaustividade, proposta por Kitchin (2013).

Volume Segundo a definição de Demchenko et al. (2013), volume é a característica mais importante e distinta em *big data*. Esta característica impõe uma série de desafios e requisitos específicos que, em geral, não são tratados pelas tecnologias tradicionais.

O ganho de volume nas bases de dados se dá, em boa parte, devido a avanços tecnológicos de instrumentação. Em diversas áreas, a tendência atual é de se coletar e armazenar dados de todos os eventos observados, de todas as atividades e sensores disponíveis, mesmo que não haja um uso previsto e direto para estes dados. Desta forma, os dados passam a ser coletados e armazenados também pelo seu uso potencial.

As características de volume em *big data* envolvem características como tamanho, escala, quantidade e dimensões. Naturalmente, esta característica apresenta um certo grau de relativismo: o que é *big* em *big data*? Áreas como física e ciências da computação apresentam características de volume muito diferentes de áreas como demografia ou saúde. Neste sentido, sugere-se que um volume adequado para classificação como *big data* seja a comparação dentro do mesmo campo de pesquisa. Por exemplo: qual é o volume de dados tipicamente coletado sobre internações hospitalares e o qual seria o volume acrescentado ao se armazenar também dados sobre todo o itinerário terapêutico, dados provenientes de sensores hospitalares, transcrições das conversas do corpo médico e do paciente e, até mesmo, dados ambientais da internação? Caso este acréscimo seja significante ao ponto de influenciar a escolha de novas tecnologias que permitam que estes dados sejam trabalhados integralmente, pode-se afirmar que trata-se de *big data*.

Velocidade Em geral, dados de *big data* são gerados em rápida velocidade. São coletados em tempo real ou em intervalos curtos de atualização. Pode-se considerar, por exemplo: dados de internação hospitalar sendo recebidos, processados e monitorados pelo Ministério da Saúde em tempo real.

Variedade Tradicionalmente, dados são coletados para uma finalidade prevista, já sendo inclusive armazenados prevendo-se um método de análise e objetivos bem definidos de seu uso. Em *big data*, são coletados todos os dados, não prevendo-se necessariamente o seu uso antecipado: todos os dados são importantes e armazenados em seu estado mais natural. Por exemplo, não se armazena necessariamente a idade de um paciente em anos, mas a data e hora de seu nascimento, permitindo que o cálculo de idade possa ser feito a qualquer hora e de forma precisa.

Desta forma, os dados podem ser do tipo "estruturados", como em planilhas e tabelas de dados, ou "não estruturados" como em textos, transcrições de falas ou sequências não lógicas de informação.

Valor Segundo Demchenko et al. (2013), esta característica é dada pelo valor agregado que os dados coletados acrescentam ao processo ou atividade. Desta forma, mesmo que a coleta de certos dados aumentem os requisitos de armazenamento e processamento, estes serão coletados caso a sua utilização seja essencial ou inovadora para o processo. Por exemplo, um sistema de informação que permita conectar dados de diversos sistemas de informação em saúde sobre um mesmo paciente, apresenta um custo considerável; contudo o valor agregado deste resultado justifica a sua execução.

Veracidade Esta característica se divide em dois aspectos: a consistência dos dados e os métodos de processamento. Enquanto o primeiro aborda a questão da veracidade dos dados, sua correspondência com a realidade, enquanto o segundo trata de questões como confiabilidade e segurança nos processos de manipulação dos dados, que permitam a sua consistência em todo o processo, obtendo-se por fim, dados verossímeis. Por exemplo, dados sobre procedimentos ambulatoriais devem ser coletados na ponta do serviço de forma fidedigna ao que foi executado de fato e o processamento destes dados não devem alterar o seu real significado.

Exaustividade *big data* pretende capturar por completo toda uma população ou sistema, onde $n = N$. Isto é, a amostra corresponde ao universo dos dados.

Em resumo, o termo *big data* pode ser aplicado quando se defronta com um grande volume de dados, produzido e atualizado em alta velocidade, com grande variedade e complexidade interna, apresentando questões específicas sobre sua veracidade, consistência e confiabilidade, tão como o valor que os dados detêm, medido pela capacidade destes gerarem novos conhecimentos e avanços (DEMCHENKO et al., 2013). Cabe aqui ressaltar que certos conjuntos de dados gerados por sistemas de informação em saúde e outras fontes nem sempre irão contemplar estas cinco características simultaneamente, mas isto não elimina a necessidade de aplicação de metodologias específicas para lidar com este tipo de conjunto de dados (HERLAND; KHOSHGOFTAAR; WALD, 2014).

Desta forma, o conceito de *big data* e sua utilização se encaixam em um campo maior, denominado *ciência de dados*. Na *ciência de dados*, além dos dados, os conceitos e métodos necessários para a coleta, análise e apresentação dos resultados são previstos.

Na área de saúde, a utilização das tecnologias de *ciência de dados* pode ser definida como:

> A habilidade de adquirir, armazenar, processar e analisar um grande volume de dados de saúde originados em várias fontes e entregar informa-

> ções significativas, permitindo a descoberta de valores e conhecimentos de maneira rápida (WANG; KUNG; BYRD, 2016).

O emprego de tecnologias de *ciência de dados* na área da saúde se dá tanto em instituições privadas, na busca de transformar a grande massa de dados produzida por planos de saúde e previdência privada em recursos para a otimização econômica de suas operações (IBM, 2012; WANG; HAJLI, 2016), quanto na área da saúde pública. Nesta última, as tecnologias de *ciência de dados* possibilitam a análise de um grande e complexo conjunto de informações de saúde de uma população, advindo de diversas escalas de análise (molecular, tecido, paciente e população), gerando hipóteses de causas e desfechos de doenças, informações para a medicina de precisão e o rastreio e predição da distribuição espacial e temporal de doenças (HERLAND; KHOSHGOFTAAR; WALD, 2014; KHOURY; IOANNIDIS, 2014).

Contudo, o emprego destas tecnologias não elimina a necessidade de observação cuidadosa do método cientifico. A produção excessiva de falsos positivos e de correlações espúrias, aliada aos vieses de seleção, escassez de variáveis de confusão e as restrições para a generalização, dentre outras questões que as tecnologias de *big data* apresentam, apenas poderão ser ultrapassadas com a combinação destas tecnologias à um conhecimento estatístico e epidemiológico sólido, multidisciplinar e enraizado em evidências (BOYD; CRAWFORD, 2012b; KHOURY; IOANNIDIS, 2014; LAZER et al., 2014; MURDOCH; DETSKY, 2013).

O emprego de tecnologias de *ciência de dados* segue um ordenamento lógico comum, um *modelo de processos de mineração de dados*. Pode-se citar o "KDD – *Knowledge Discovery in Databases*" (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996a; HERLAND; KHOSHGOFTAAR; WALD, 2014), "SEMMA – *Sample, Explore, Modify, Model, Assess*" (SAS, 2016) e "CRISP-DM – Cross-Industry Standard Process for Data Mining" (CHAPMAN et al., 2000). Em comum, estas três metodologias englobam etapas relativas a aquisição, armazenamento, recuperação de dados, mineração de informações e por fim, análise e visualização dos dados. Estes processos, suas origens e aplicações, serão explorados em detalhe no capítulo Metodologia (p. 40).

No contexto de aplicação destes modelos de processos de mineração de dados na saúde pública, a etapa de aquisição compreende a coleta de informações secundárias junto a repositórios de dados públicos ou privados ou a coleta primária de dados através de pesquisas e inquéritos, tão como a coleta de dados não estruturados, como textos, opiniões e comentários, disponíveis em redes sociais, blogs, jornais e outras mídias. As etapas de armazenamento e recuperação de dados envolvem tecnologias específicas que permitem lidar com grandes massas de dados sem comprometer a velocidade das operações de leitura e apresentação dos arquivos. A etapa de mineração de dados envolve processos para a detecção de padrões, associações e tendências temporais e espaciais, envolvendo em geral

algoritmos autônomos e autorregulados para a realização destas tarefas. A última etapa, referente a análise e visualização de dados, visa a extração e apresentação de informações permitindo ao utilizador a descoberta de novas relações.

Nesta última etapa reside uma importante relação entre a tecnologia de *ciência de dados* e usuários: a descoberta e comunicação da informação. Neste passo, a utilização de tecnologias de *ciência de dados* ganha significado prático ao usuário, permitindo que um grande e amorfo conjunto de dados seja visualizado e compreendido de maneira interativa, atribuindo significado aos dados (ENDERT; BRADEL; NORTH, 2013; HUANG et al., 2015). Assim, a análise de dados através das tecnologias de *ciência de dados* deve permitir informar aos gestores sobre diferentes possibilidades de políticas públicas e na apresentação das informações para o público em geral (FUNG; TSE; FU, 2015).

Retornando às eras da evolução da epidemiologia, proposta por Susser e Susser (1996a), Susser e Susser (1996b), compreende-se que *ciência de dados* e *big data* podem configurar um novo paradigma. Conforme afirma Boyd e Crawford (2012a):

> *Big Data creates a radical shift in how we think about research... [It offers] a profound change at the levels of epistemology and ethics. Big Data reframes key questions about the constitution of knowledge, the processes of research, how we should engage with information, and the nature and the categorization of reality... Big Data stakes out new terrains of objects, methods of knowing, and definitions of social life (BOYD; CRAWFORD, 2012a).*

Segundo Kitchin (2014), a análise de *big data* possibilita uma abordagem epistemológica completamente nova para a compreensão do mundo. Ao invés de testar uma teoria analisando os dados pertinentes, apresentam-se novos métodos para buscar ideias e intuições nascidas à partir dos dados.

Desta forma, sobre a produção de conhecimento, argumenta-se que que *big data* apresenta a possibilidade de um novo paradigma para diversas disciplinas (KITCHIN, 2014).

Conforme o quadro 2, Hey, Tansley e Tolle (2009) sugerem um quarto paradigma da ciência, que se baseia na contínua disponibilidade de *big data* e novos métodos de análise.

Conforme Kuhn (1970), um paradigma constitui um modo aceito de questionar o mundo e sintetizar o conhecimento de uma proporção substancial de pesquisadores de uma disciplina em um momento no tempo. Desta forma, segundo o autor, periodicamente emergem novos modos de pensar que desafiam as teorias e abordagens previamente aceitas por essa comunidade, ao passo que a ciência dominante não consegue mais responder a todas as perguntas e requer a formulação de novas ideias (KITCHIN, 2014). De modo diferente, Boyd e Crawford (2012a) propõe a transição dos paradigmas da ciência segundo

Quadro 2 – Os quatro paradigmas da ciência

Paradigma	Natureza	Forma	Momento
Primeiro	Ciência experimental	Empirismo, descrição dos fenômenos naturais.	Pré-renassença
Segundo	Ciência teórica	Modelagem e generalização.	Pré-computadores
Terceiro	Ciência computacional	Simulação de fenômenos complexos.	Pré-*big data*
Quarto	Ciência exploratória	Dependente de dados, exploração estatística e *data mining*.	Momento atual

Fonte: Compilado a partir de Hey, Tansley e Tolle (2009).

mudanças fundamentais nos dados disponíveis e nos métodos de análise, propondo que a ciência adentra agora em um quarto paradigma, baseado na crescente disponibilidade de *big data* e novas formas de análise.

Ainda que a definição de paradigma de Kuhn possa ser criticada (WALKER, 2010), esta noção apresenta-se útil para conduzir os debates sobre *big data* e suas mudanças na produção de conhecimento, considerando-se que boa parte das afirmações sobre *big data* estipulam o surgimento de um novo paradigma, ainda que a forma que esta nova epistemologia toma seja contestada (BOYD; CRAWFORD, 2012a).

Duas correntes epistemológicas se apresentam frente a *ciência de dados*, a de uma nova era do empirismo e a de *data driven science* (PROVOST; FAWCETT, 2013). Na primeira corrente, conforme argumenta Anderson (2008), a correlação supera a causa, a ciência pode avançar mesmo sem modelo coerentes ou teorias unificadoras. Nesta corrente, a mineração de conhecimento em *big data* pode revelar relações e padrões que sequer sua existência era conhecida e, logo, não poderiam ser averiguados pelos métodos tradicionais de teoria primeiro e corroboração em seguida. Conforme Siegel (2013), em geral não se sabe sobre os mecanismos de causa e, em geral, não é necessário conhecimento sobre ele. O objetivo maior é predizer, não é compreender a realidade.

Este modo de se fazer ciência se apresenta de forma puramente indutiva em sua natureza. Considera-se que os dados – puramente – são capazes de traduzir a realidade com os atuais métodos de pesquisa. A afirmação de Sextus Empiricus (epígrafe desde capítulo) estabelece que método científico estaria "lidando com o impossível, posto que as partículas são infinitas e indefinidas". Esta afirmação pode estar possivelmente superada pelos novos métodos de coleta e análise de dados.

Contudo, ainda que o *big data* busque ser exaustivo, capaz de assimilar as infinitas e indefinitas partículas da realidade, ele será sempre uma representação ou uma amostra da realidade, conformada pela tecnologia, ontologia e ambiente regulatório, sempre sujeito ao erro amostral (CRAWFORD, 2013; KITCHIN, 2013).

A corrente de *data driven science* busca ser mais aberta a raciocínios híbridos de

dedução e indução, ainda que se estabeleça nas raízes mais firmes do método científico tradicional. Neste sentido, esta corrente busca gerar hipóteses também a partir dos dados e não somente à partir da teoria (MAASS et al., 2018; PROVOST; FAWCETT, 2013; KELLING et al., 2009). A proposta epistemológica desta corrente é buscar hipóteses nos dados com os métodos apropriados para *big data* e comprová-los através dos métodos tradicionais de verificação de hipóteses (KITCHIN, 2014).

Espera-se que esta corrente forme o novo paradigma da ciência na era da *ciência de dados* pois sua epistemologia favorece a extração adicional de conhecimento a partir dos dados sem descumprir o rigor científico aceito. Uma ciência puramente dedutiva negligenciaria os avanços tecnológicos e metodológicos conquistados, que permitem compreender conjuntos muito maiores de dados, de forma interligada e dinâmica, de formas que seriam antes impossíveis de trabalhar (KELLING et al., 2009; LOUKIDES, 2010; MILLER, 2010).

Nesta linha, dentre os diversos campos que podem se beneficiar deste novo paradigma epistemológico, encontra-se a epidemiologia. A epidemiologia tradicional é *theory driven*: ela parte de um paradigma que requer a compreensão completa dos fatores associados a um desfecho e formulação de uma teoria clínica sólida, para assim justificar e viabilizar o estudo de um modelo empírico da realidade. As possibilidade da *ciência de dados* expandem a epidemiologia para uma abordagem *data driven*: os cinco "V" dos dados possibilitam a formulação de hipóteses à partir dos dados, que podem ser posteriormente verificadas por métodos mais tradicionais ou de mineração de dados.

Nesta nova configuração, a saúde pública se beneficia de um método misto, buscando na abordagem indutiva da *ciência de dados* a construção de novas hipóteses sem necessitar de apoio ou confirmações clínicas *a priori*, e na abordagem dedutiva para a verificação destas hipóteses com o rigor científico necessário para a construção de políticas públicas.

3 Metodologia

> *Computadores nos prometeram uma fonte de conhecimento mas entregaram uma enchente de dados.*
>
> —Frawley, Piatetsky-Shapiro e Matheus (1992).

O esforço em padronizar e documentar a análise de dados é tão antigo quanto a sua própria prática. Conforme as necessidades particulares de cada campo, padrões são estabelecidos e são revisados para dar conta dos diversos processos envolvidos, indo deste a definição do problema de analise e aquisição dos dados até a divulgação de resultados e geração de conhecimento. Rotondo e Quilligan (2020) aponta que um grande desafio para analistas de dados no século XXI é o desenvolvimento e disseminação de padrões que sejam amplamente aceitos em vários campos e áreas.

No século XX, alguns *standards* foram criados para a análise de dados, oriundos de setores ligados às indústrias e tecnologia. Gradativamente, estes padrões e modelos de processamento *process models* são utilizados e adaptados por outras áreas, incluindo a pesquisa acadêmica.

Pode-se citar três grandes modelos de processamento estabelecidos e reconhecidos: *Knowledge Discovery in Databases* (KDD), *Sample, Explore, Modify, Model, Assess* (SEMMA) e o *Cross-Industry Standard Process for Data Mining* (CRISP-DM). Os três modelos são detalhados nas seções a seguir, procurando aplicar os seus passos à saúde pública brasileira.

3.1 KDD

O processo denominado *Knowledge Discovery in Databases* (KDD) se originou no início da década de 1990, durante o avanço e disseminação de computadores em diversos setores da sociedade, nas ciências e mercado. Frente a velocidade do crescimento e variedade das bases de dados, os métodos estatísticos tradicionais – de análise manual – se apresentavam insatisfatórios e um novo método de abordagem era necessário para preencher o hiato entre a produção de dados e a produção de conhecimento (PIATETSKY-SHAPIRO, 1990).

Na década de 1990, alguns métodos para a detecção de padrões em dados já existiam, ainda que apresentassem problemas frente ao volume ou complexidade dos dados. O conjunto de métodos de detecção de padrões em dados receberam diversos nomes, como "*data mining*", "*knowledge extraction*", "*information discovery*", "*information harvesting*", "*data archeology*" e "*data pattern processing*" (FAYYAD; PIATETSKY-SHAPIRO; SMYTH,

1996b). Contudo, era latente a dificuldade de emprego direto destes métodos para obtenção de resultados úteis: etapas anteriores e posteriores a ele eram necessárias.

Acompanhando esta necessidade, um *workshop* durante o *International Joint Conference on Artificial Intelligence* de 1989 (Detroit) foi conduzido, visando enfatizar que "conhecimento" era o produto final da descoberta, cunhando a expressão *Knowledge Discovery in Databases* (KDD) (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996b).

Após o *workshop*, alguns *papers* foram publicados ao longo da mesma década, visando formalizar o processo como um método e popularizar o conceito nas comunidade de Inteligência Artificial e Aprendizado de Máquina (PIATETSKY-SHAPIRO, 2000).

A diferenciação entre KDD e *data mining* se encontra na formalização dos processos anteriores e posteriores necessários para a melhor extração de conhecimento das bases de dados. Enquanto KDD se refere ao processo global de extração de conhecimento útil das bases de dados, *data mining* se refere a um dos passos deste processo (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996b).

De modo abstrato, KDD é um campo que se ocupa com o desenvolvimento de métodos e técnicas que possibilitem gerar uma compreensão dos dados disponíveis. Em geral, o processo constitui em transformar dados brutos (tipicamente muito volumosos para se compreender e digerir diretamente) em outras formas que possam ser mais compactas (relatórios, por exemplo), mais abstratas (um modelo geral dos dados) ou mais útil (um modelo preditivo para estimação de casos futuros (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996a).

Mais formalmente, o processo KDD é definido como a "extração não trivial de conhecimento implícito, previamente desconhecido e potencialmente útil dos dados"(FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996c; FRAWLEY; PIATETSKY-SHAPIRO; MATHEUS, 1992).

Nesta definição, *dados* se referem ao conjunto de fatos representados em uma *base de dados*. Denomina-se *processo* por ser composto de uma série de passos previstos e necessários, que envolvem a preparação dos dados, busca por padrões, avaliação do novo conhecimento obtido e refinamentos, conduzidos de forma iterativa e interativa (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996b).

Conforme define Fayyad, Piatetsky-Shapiro e Smyth (1996b),

> *KDD Process is the process of using the database along with any required selection, preprocessing, subsampling, and transformations of it; to apply data mining methods (algorithms) to enumerate patterns from it; and to evaluate the products of data mnining to identify the subset of the enumerated patterns deemed "knowledge".*

Desta forma, os passos básicos do processo KDD são (FAYYAD; PIATETSKY-

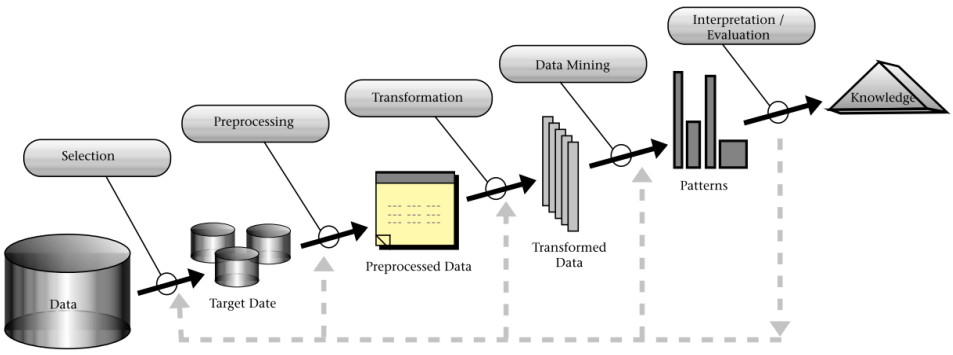
SHAPIRO; SMYTH, 1996b):

1. Desenvolver uma compreensão de domínio e levantamento de conhecimento prévio relevante, identificando o objetivo do processo KDD;
2. Criar um conjunto de dados alvo, o que significa selecionar um conjunto de dados ou desenvolver uma amostra de registros ou um subconjunto de variáveis que possibilitará a descoberta de conhecimento;
3. Redução da dimensionalidade dos dados usando metodologias próprias para reduzir o número de variáveis sem prejuízo da variabilidade e complexidade dos dados;
4. Escolher um método de *data mining* alinhado ao objetivo levantado no passo 1, como sumarização, classificação, regressão,*clustering* e outros;
5. Definir o algoritmo específico de um dos métodos de *data mining*, prevendo quais parâmetros os modelos usarão;
6. Aplicação do algoritmo de *data mining* para a busca de padrões;
7. Interpretação dos padrões obtidos, em geral através de visualização;
8. Consolidação do conhecimento gerado, incorporando-o em sistemas que podem se beneficiar deste novo conhecimento ou documentado-o para publicizar os achados.

O processo KDD pode envolver uma significante iteração, contendo *loops* entre passos até que se obtenha uma resposta adequada para o próximo passo.

De forma geral, o processo KDD pode ser resumido em 5 passos básicos, conforme a figura 2. Estes passos básicos serão detalhados nas seções a seguir, aplicando-os ao contexto da Saúde Pública no Brasil.

Figura 2 – O processo KDD



Fonte: Fayyad, Piatetsky-Shapiro e Smyth (1996a)

Coleta e seleção de dados

A Saúde Pública brasileira dispõe hoje de vasta gama de bases de dados, seja através de seus Sistemas de Informação em Saúde ou dados sócio-demográficos. Após o processo de redemocratização da república no final da década de 1980 e o estabelecimento e consolidação do Sistema Único de Saúde (SUS) nas décadas seguintes, a coleta e produção de dados do SUS foi gradativamente reforçada, culminando com a criação do Departamento de Informática do SUS (DataSUS) em 1991.

Atualmente, o DataSUS é encarregado pela alimentação, manutenção e disseminação de dados provenientes de diversos Sistemas de Informação em Saúde, onde pode-se citar o Sistema Nacional de Mortalidade (SIM), Sistema de Informações de Nascidos Vivos (SINASC), Sistema de Informações Hospitalares (SIH), Sistema de Informações Ambulatoriais (SIA), Sistema de Informações de Atenção Básica (SIAB), Sistema Nacional de Agravos de Notificação (SINAN), Registro de Eventos em Saúde Pública (RESP), dentre vários outros. Além destes, outros sistemas e subsistemas reúnem dados específicos, como o SISMAMA, SISCOLO, SISAGUA e Sistemas de Vigilância Epidemiológica (SIVEP) (ALMEIDA; ALENCAR, 2000; MARIN, 2010).

Cada um destes sistemas de informação apresentam histórias únicas sobre sua criação, atendendo a demandas específicas para a administração pública, o que pode explicar – ainda que parcialmente – a complexidade interna destes sistemas e a aparente incomunicabilidade entre eles.

De forma geral, o ciclo de vida é coberto por diferentes sistemas, de modo solitário ou sobrepostos em alguns momentos. Por exemplo, o SIAB apresenta dados sobre o processo de planejamento familiar e acompanhamento da gestação e pré-natal; o SINASC reúne dados sobre o nascimento; sistemas como o SIH, SIA e SINAN reúnem dados sobre o processo saúde-doença da infância, idade adulta e avançada; e por fim o SIM reúne os dados sobre o óbito.

Alguns destes sistemas de informação em saúde surgiram com a missão de acompanhamento epidemiológico da população, como o SINAN, enquanto outros são advindos de informações cartoriais sobre mortalidade e nascimento (SIM e SINASC) ou concebidos inicialmente para fins gerenciais financeiros como o SIH e SIA. Independente de sua origem, estes sistemas podem ser usados para fins epidemiológicos desde que se leve em consideração seu viés de origem e limitações nas análises, controlando assim as inferências para o escopo de cada sistema (BITTENCOURT; CAMACHO; LEAL, 2006). A tabela 1 apresenta dados sobre o volume de alguns destes sistemas.

Além dos sistemas de informações em saúde, o Saúde Pública se beneficia diretamente de levantamentos e inquéritos sócio-demográficos, em geral conduzidos pelo Instituto Brasileiro de Geografia e Estatística (IBGE) (VIACAVA, 2002). Pode-se citar, dentre

Tabela 1 – Volume de alguns sistemas de informação de saúde nacionais

Sistema	Período	Registros
SIA	1994 a 2017	63.987.334.421
SIH	1984 a 2018	403.452.248
SINASC	1994 a 2017	71.270.156
CNES	2005 a 2018	38.223.272
SIM	1979 a 2017	37.372.780

Fonte: compilado pelo autor com dados do DataSUS.

outras pesquisas, o Censo Demográfico, a Contagem Populacional, a Pesquisa Nacional de Amostra de Domicílios (PNAD), a Pesquisa de Orçamentos Familiares (POF) e a Pesquisa Nacional de Saúde (PNS).

Estas pesquisas reúnem dados gerais sobre a população, trabalho e emprego, apresentando informações de contexto indispensáveis para o estudo da saúde na população.

Além de pesquisas sobre a população, importantes fontes de dados são encontradas nos órgãos de monitoramento ambiental, climático e de recursos físicos. Por exemplo, dados sobre desmatamento e focos de incêndios do Instituto Nacional de Pesquisas Espaciais (INPE), dados sobre precipitação e temperatura do Instituto Nacional de Meteorologia, dados sobre mananciais e abastecimento de água da Agência Nacional das Águas e dados de recursos minerais, solos e atividades extração mineral da Agência Nacional de Mineração (ANM)[1] e dados sobre desastres coletados pelas Secretarias de Estado de Meio Ambiente e pela Defesa Civil. Todas estas fontes apresentam dados essenciais para o estudo dos processos de saúde-doença.

As fontes de dados até o momento apresentadas podem ser categorizadas como de "dados estruturados". Neste formato, os dados são apresentados – em geral – de forma tabular, com registros de casos nas linhas e variáveis e parâmetros de acompanhamento nas colunas. Este é o formato mais tradicional, utilizado pela estatística convencional.

Contudo, dados do tipo "não estruturados" surgem como potenciais fontes de informação para a Saúde Pública. Por exemplo, transcrições de consultas, anamneses, prontuários eletrônicos e postagens em redes sociais contêm textos e outros tipos de dados que documentam etapas do atendimento médico onde os dados são escassos e do estilo de vida dos usuários, suas opiniões e atividades mais frugais. Estes dados podem ser analisados de forma semi-automática ou automática e em conjunto com os dados estruturados dos sistemas de informação em saúde convencionais (CHOWKWANYUN, 2019).

Diante da vasta gama de possibilidade de fontes de dados, se faz necessário considerar suas características frente aos objetivos de estudo. Os sistemas de informação em saúde costumam disponibilizar dados brutos, além de agregações por município e anos. Nos dados brutos, cada linha é separadamente um evento – um nascimento, um óbito ou

[1] Antigo Departamento Nacional de Produção Mineral (DNPM)

uma internação hospitalar, por exemplo. Através dos dados brutos é possível reproduzir as agregações disponibilizadas pelo Ministério da Saúde mas também realizar outras agregações, com mais opções espaciais e temporais, tão como utilizar uma gama maior de opções para filtros de seleção.

A vantagem da maior gama de informações sobre os eventos apresenta uma consequência: os dados brutos costumam ser mais complexos de trabalhar, apresentam uma linguagem mais técnica e específica, as documentações podem ser deficientes, o volume de dados pode demandar soluções computacionais específicas e, em geral, não permitem o cruzamento entre as bases de dados. Os microdados públicos divulgados pelo DataSUS para os sistemas de informação em saúde são anonimizados por razões de privacidade e segurança.

Desta forma, apesar dos sistemas de informações em saúde apresentarem dados sobre o ciclo de vida por completo, não é possível atualmente percorrer o itinerário terapêutico de um paciente, por exemplo, no caso de um nascimento pré-maturo, procedimentos ambulatoriais, internação e óbito. Cabe mencionar que existem parcerias diretas com o Ministério da Saúde que visam possibilitar estas análises em ambiente controlado, sob vista de comitês de ética em pesquisa, como o "Centro de Integração de Dados e Conhecimento para Saúde" (CIDACS) da Fundação Oswaldo Cruz. Pode-se citar também os esforços do Laboratório Nacional de Computação Científica (LNCC), através do *Data Extreme Lab* (DEXL Lab), para pesquisa e operacionalização da aplicação de métodos de *ciência de dados* e *big data* em diversos campos de aplicação.

Ainda que não seja convencional o cruzamento de dados entre os sistemas de informação, os dados que cada um possuem isoladamente oferecem grande potencial para pesquisa científica, apresentando gradativamente no tempo, melhor qualidade de preenchimento e representatividade no território nacional.

Frente a diversidade de fontes de dados disponíveis, encontram-se questões sobre a qualidade dos dados, um fator que impacta diretamente sua usabilidade. Na seção seguinte, esta questão será abordada.

Pré-processamento

A origem de cada sistema de informação em saúde, sua história de manutenção e diversidade de usos, tão como características intrínsecas de outras pesquisas como Censo Demográfico, caracterizam os dados disponíveis com uma gama de especifidades que precisam ser levadas em conta antes de qualquer análise (JORGE; LAURENTI; GOTLIEB, 2007).

Esta etapa de pré-processamento visa conduzir inicialmente uma análise exploratória dos dados, com o objetivo de qualificar a qualidade geral dos dados selecionados para

análise. Com os resultados desta qualificação, algumas ações pode ser consideradas, como eliminação de registros de má qualidade ou imputação de valores faltantes.

A seguir, algumas destas especificidades do âmbito dos sistemas de informação em saúde serão discutidas.

Cobertura

A implantação dos sistemas de informação em saúde no território brasileiro se deu de forma gradual e independente, de forma relacionada ao princípio de descentralização político-administrativa (BRANCO, 1996; ALMEIDA, 1998). Alguns sistemas de informação já eram existentes previamente à estruturação do SUS, enquanto outros foram criados após sua existência.

Desta forma, a cobertura destes sistemas, em seu início, costuma estar ligada às capitais estaduais e grandes núcleos urbanos, se propagando gradativamente para o interior dos estados. Grandes avanços foram feitos nas últimas décadas, principalmente ao se considerar os desafios continentais para coleta de dados e centralização no território brasileiro. Pode-se citar, por exemplo, os avanços obtidos com a informatização do processo de coleta e envio de dados via Internet para o Ministério da Saúde (MS).

Atualmente, a cobertura dos sistemas de informação em saúde como SIM, SINASC, SIH e SIA, atinge todos os municípios brasileiros. Contudo, outros sistemas de informação em saúde tem cobertura mais dispersa, como os de vigilância epidemiológica. Esforços do Ministério da Saúde tem sido realizados em torno do "e-SUS", uma estratégia para "desenvolver, reestruturar e garantir a integração desses sistemas, de modo a permitir um registro da situação de saúde individualizado por meio do Cartão Nacional de Saúde" (Ministério da Saúde, 2019).

Qualidade do preenchimento

A qualidade do preenchimento dos sistemas de informação em saúde pode ser entendida em termos da existência/inexistência da informação (dados faltantes) e a qualidade propriamente dita da informação registrada (CORREIA; PADILHA; VASCONCELOS, 2014; LIMA et al., 2009).

Atualmente, informações essenciais dos sistemas, como datas e diagnósticos, estão sempre presentes (De Mathias; De Soboll, 1998). Contudo, questionamentos ainda persistem sobre a correta diferenciação entre causa principal e secundária, o que leva potencialmente a geração de "*garbage codes*" em campos como causa básica do óbito ou internação (AGUIAR et al., 2013).

Campos como cor/raça, escolaridade e ocupação apresentam, em geral, má qualidade (BRAZ et al., 2014); principalmente pelo não preenchimento destes campos. Melhorias

nestes campos demandam maiores esforços dos sistemas de saúde municipais e estaduais em sua coleta, crítica e acompanhamento contínuo.

Melhorias foram obtidas com a informatização da coleta e processos de crítica automática dos dados, apontando incongruências no momento da digitação, em geral impedindo o envio do registro até a questão ser solucionada.

Comparabilidade

A extensão temporal e territorial dos sistemas de informação em saúde levam a questões específicas sobre a comparabilidade dos dados. Por exemplo, a implementação da 10a edição da Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde (CID-10) em 1989 e posteriores revisões em 2003 e 2008, criam a necessidade de cuidados para a comparação de informações sobre diagnóstico.

A evolução da malha territorial brasileira de municípios também leva a desafios técnicos de comparação, com a criação, extinção e agregação de municípios. A tabela 2 ilustra em parte este desafio, também denominado de "*Modifiable Areal Unit Problem* (MAUP) (FOTHERINGHAM; ROGERSON, 2009). Metodologias como redistribuição de casos entre municípios e criação de áreas mínimas comparáveis para os municípios podem ser adotadas (EHRL, 2017).

Tabela 2 – Número de municípios nos censos demográficos, segundo as grandes regiões brasileiras

Região	Censo 1980	Censo 2010
Região Norte	203	449
Região Nordeste	1.375	1.794
Região Sudeste	1.410	1.668
Região Sul	719	1.188
Região Centro-Oeste	284	466
Brasil	3.991	5.565

Fonte: dados compilados pelo autor.

Sobre o SINASC, especificamente, pode-se também citar a mudança na coleta da informação de raça/cor, sendo ora referente à mãe e ora ao nascido vivo (PEDRAZA, 2012; OLIVEIRA et al., 2015).

Atualização e disponibilidade dos dados

Os sistemas de informação em saúde apresentam diferentes periodicidades de atualização. De forma geral, sistemas de origem epidemiológica como o SIM e SINASC divulgam novos dados anualmente, com defasagem de um a dois anos. Tal defasagem se faz necessária para apuração completa de todos os registros, verificação de qualidade e, quando necessário, busca ativa de casos e demais confirmações. Já os sistemas de origem administrativa, como SIH e SIA, apresentam atualização mensal, com defasagem de um a

dois meses em média, ainda que estados da região norte e nordeste apresentem atrasos maiores.

Os dados dos sistemas de informação em saúde são disponibilizados pelo DataSUS de forma agregada através da interface TabNET, que consiste em um conjunto de páginas na Internet que permite ao usuário a escolha do sistema de informação, seleção de indicadores, filtros diversos sobre os dados e seleção do nível de agregação espacial (regiões, unidades federativas ou municípios) e agregação temporal (anual ou mensal, de acordo com o sistema de informação em saúde).

O DataSUS disponibiliza os dados brutos (microdados) através de um endereço de transferência de arquivos na Internet (FTP). Estes arquivos são oferecidos no formato proprietário "DBC", podendo ser lido através de um aplicativo desenvolvido pelo próprio DataSUS denominado "TabWIN" (plataforma Windows) ou através de bibliotecas especificas de leitura de dados, como a "read.dbc" (PETRUZALEK, 2016).

Documentação

Em geral, a documentação dos sistemas de informação em saúde oferecida pelo DataSUS é fornecida por meio de um dicionário de dados contendo o nome dos campos e breve descrição de seu conteúdo.

Avanços neste sentido são necessários, como descrição mais completa das variáveis, códigos de rótulos das variáveis categóricas, documentação do histórico de versões de dados, entre outros.

Transformação

Os dados pré-processados necessitam ser armazenados de modo a possibilitar a execução das etapas seguintes de mineração e visualização. Para tanto, estratégias específicas necessitam ser adotadas para oferecer uma rápida localização e recuperação dos dados.

Os dados advindos dos sistemas de informações de saúde podem ser armazenados através de bancos de dados relacionais tradicionais, como o PostgreSQL. Contudo, questões de performance devem ser observadas, podendo ser consideradas soluções do tipo "NoSQL" (HAN et al., 2011).

Nesta etapa, métodos de redução de dimensionalidade também podem ser empregados para a eliminação de variáveis redundantes através da criação de novas variáveis (fatores) de resumo.

Para o desenvolvimento desta etapa, se faz necessária a adoção de uma estratégia que permita a otimização dos recursos computacionais disponíveis para o armazenamento

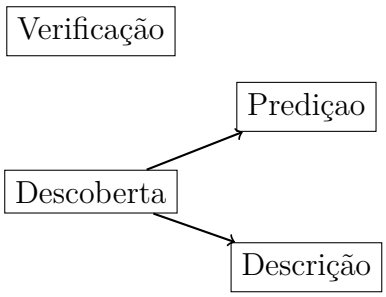
e recuperação de informações, como através da metodologia "MapReduce" (MAITREY; JHA, 2015).

3.1.1 Mineração de dados

Mineração de dados consiste na aplicação de algoritmos específicos para a verificação de padrões nos dados. Em geral, esta etapa envolve uma série de passos repetidos e iterativos de acordo com o algoritmo escolhido (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996a).

Estes algoritmos cumprem objetivos que podem ser esquematizados como na figura 3.

Figura 3 – Objetivos de *data mining*



Fonte: elaborado pelo autor.

Objetivos de *verificação* consistem em verificar hipóteses previamente formuladas, já as etapas de *descoberta* não possuem hipóteses *a priori*, podendo serem formuladas pelo algoritmo de forma autônoma.

Neste último caso, os algoritmos podem ser do tipo de *predição*, onde valores futuros são estimados; ou de *descrição*, quando o sistema apresenta padrões nos dados de forma legível para humanos (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996b). Desta forma, os algoritmos de *data mining* consistem em ajustar modelos ou detectar padrões à partir dos dados observados.

Boa parte dos métodos de *data mining* consistem em técnicas de aprendizado de máquina (*machine learning*), reconhecimento de padrões e da estatística: classificação, regressão, clusterização e sumarização, dentre outros (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996b). Para cada um destas técnicas, diversos algoritmos estão disponíveis, atendendo a diferentes requisitos específicos de aplicabilidade e tipos de resultados desejados.

Classificação: aprender uma função que mapeia (classifica) os dados em classes pré-definidas.

Regressão: aprender uma função que reflete o comportamento geral dos dados, capaz de predizer valores e descobrir relações funcionais entre as variáveis.

Clusterização: identificar um conjunto finito de categorias ou *clusters* para descrever os dados.

Sumarização: criar uma descrição compacta dos dados, como sumários, regras de associação e recursos de visualização multivariados.

Interessante notar que parte destas técnicas já fazem parte do ferramental estatístico, como análise de *clusters*, análise fatorial, séries temporais e análises espaciais. A inovação desta etapa reside no emprego destas técnicas em um grande volume de dados, enfrentando questões acerca da aplicabilidade de conceitos de inferências estatística e testes de hipóteses (FRANKE et al., 2016; LIN; LUCAS, 2013).

3.1.2 Visualização

Aliada aos processos estatísticos de mineração de dados, a visualização permite ao pesquisador uma percepção geral sobre o comportamento e distribuição dos dados. O objetivo das representações gráficas é resumir os dados e enfatizar suas principais características (TUKEY, 1977).

A Análise Exploratória de Dados aplicada a *big data* necessita de adaptações para lidar com as características de volume e velocidade, dentre outras. A criação de painéis de visualização tem sido um recurso utilizado amplamente para a apresentação e representação de dados, permitindo o resumo dos mesmos através de gráficos especializados e interativos (ENDERT; BRADEL; NORTH, 2013; PUTS; DAAS; WAAL, 2015). O interesse na visualização de dados no contexto de *big data* tem crescido devido a habilidade humana de detectar rapidamente padrões e relações (CHOO; PARK, 2013).

Painéis de controle (*dashboards*) são recursos de visualização de dados que permitem o monitoramento da situação através da apresentação simultânea de gráficos, estatísticas e índices em um mesmo recorte, permitindo ao leitor a recepção simultânea de diferentes informações. A utilização de painéis de controle teve início na década de 1980, no mercado privado e buscava, neste contexto, a apresentação de informações indispensáveis ao executivo sobre a performance da organização para a tomada de decisões mais rápidas (ECKERSON, 2006; FEW, 2006a; FEW, 2006b; KERZNER, 2013).

Aplicando a lógica de painéis de controle na saúde pública, seu objetivo passa ser o de visualizar e analisar a performance do sistema de saúde ou de um recorte dele, traduzindo as políticas públicas de saúde em métricas e metas. Este acompanhamento possibilita o ajuste fino das ações para a perseguição dos objetivos, disseminando a situação atual e os objetivos pretendidos (ECKERSON, 2006).

A utilização de painéis de controle requerem uma infraestrutura de integração de dados, como a que é permitida através das tecnologias de *big data*, levando ao desenvolvimento de três ações básicas: monitoramento, análise e gerenciamento. O monitoramento diz respeito a comparação das métricas do sistema com as metas pretendidas, posicionando a situação atual entre valores mínimos e máximos de operação regular. A análise é o processo de compreender as razões determinantes das métricas obtidas e as possíveis anomalias que podem estar ocorrendo. Por fim, o gerenciamento é o processo de reação frente o monitoramento e análise, visando modificar a realidade do sistema.

Compreendendo estes três processos básicos, um painel de controle costuma apresentar três camadas de interação: visão gráfica resumida, visão multidimensional e a visão detalhada. Na visão gráfica resumida, são exibidas estatísticas e indicadores chaves para o resumo da situação, em geral de forma gráfica e comparados a limites de operação aceitáveis. Na visão multidimensional, os dados geradores das estatísticas e indicadores são exibidos de maneira longitudinal e espacial, permitindo a agregação e desagregação dos dados. Já na visão detalhada, os dados são exibidos através de registros desagregados. Esta hierarquia de camadas de apresentação visa permitir ao usuário o monitoramento da situação, a investigação de suas causas através do acompanhamento no tempo e espaço do desempenho das variáveis e indicadores e, por fim, a observação detalhada da origem da situação nos dados originais (ECKERSON, 2006; FEW, 2006b).

A utilização de painéis de controle na saúde vem se popularizando, podendo ser encontrado no monitoramento clínico de pacientes (DOWDING et al., 2015), no acompanhamento administrativo de instituições privadas de saúde (BASKETT; LEROUGE; TREMBLAY, 2008; HARRINGTON et al., 2006), tão como tendo diversas aplicações na saúde pública.

Nesta, dentre outros empregos, encontra-se utilizações de painéis de controle para o monitoramento do sistema de saúde (JINPON; JAROENSUTASINEE; JAROENSUTASINEE, 2011; KOSTKOVA, 2013; KOSTKOVA et al., 2014; RANA, 2015; TURNER, 2009; WEIR et al., 2009; YI et al., 2008). Deste tipo de utilização, pode-se destacar uma proposição de monitoramento do sistema de saúde brasileiro (VIACAVA et al., 2004), servindo-se de proposições da OPAS para o acompanhamento e melhoria da performance de sistemas de saúde nas Américas (PAHO, 2011).

Também na área da saúde pública, são encontradas utilizações específicas para o monitoramento de doenças, se destacando nas últimas décadas a aplicação no monitoramento do vírus Influenza (CHENG et al., 2011; FAN et al., 2010; PODGORNIK et al., 2007).

No acompanhamento de doenças através de painéis de controle, destaca-se a tendência de apresentação dos dados em três formas básicas nas páginas iniciais: textos, mapas e gráficos. Os textos são apresentados na forma de tópico e frases, resumindo a

situação da doença de interesse, como por exemplo: “estável” e “moderada”. Os mapas apresentam a distribuição espacial de um indicador em um determinado período, e enquanto os gráficos apresentam a variabilidade temporal de um indicador para todas as regiões, confrontando a performance do indicador com parâmetros máximos e mínimos considerados regulares.

As camadas seguintes de apresentação possibilitam ao usuário a personalização dos relatórios, através da seleção de áreas geográficas e períodos de interesse, onde destaca-se também a opção de exportação dos dados para formatos de escolha do usuário.

Torna-se importante ressaltar que o processo de comunicação e saúde realizado através destes painéis não pode se limitar à transmissão de informações de saúde para a comunidade científica e população. A articulação de saberes dos campos da saúde e da comunicação deve promover a relação entre discurso e mudança social, resultando numa mudança comportamental baseada em evidências, acolhendo as contribuições da diversidade de atores envolvidos no processo (ARAÚJO; CARDOSO, 2007).

3.1.3 Avanços e variações da metodologia KDD

A metodologia KDD, estabelecida na década de 1990, segue sendo utilizada em diversos setores e recebendo atualizações e revisões. Pode-se destacar a revisão proposta por Collier et al. (1998), que introduz o conceito de re-iteração ao processo e a integração do método ao campo de engenharia de ontologias (GOTTGTROY, 2007).

KDD como um ciclo contínuo

A introdução de re-iterações ao método KDD estabelece um lógica de exame contínuo de seus resultados e adequação, se necessária, de seus passos ao objetivo inicial (ROTONDO; QUILLIGAN, 2020), conforme ilustrado pela figura 4.

Collier et al. (1998), através do *Center for Data Insight* extendo o método KDD de Fayyad, Piatetsky-Shapiro e Smyth (1996a) com outras propriedades:

Definições de questões: Compreende que a mineração de dados não é uma panacéia metodológica capaz de, automaticamente, responder toda e qualquer questão. Assim, se faz necessário definições precisas das questões e objetivos a serem respondidos.

Aproveitamento dos resultados: O método KDD tradicional se encerra na avaliação dos resultados produzidos, não fornecendo um encaminhamento sobre a utilização destes resultados no processo de decisão final. A revisão propõe a necessidade de apontar resultados que possam ser colocados em prática.

Figura 4 – O método KDD como um processo iterativo

Fonte: Rotondo e Quilligan (2020)

Iteração: Ainda que o método KDD tradicional já considera o retorno a fases prévias para melhoria do resultado final, esta característica necessita ser mais implícita ao modelo.

Ontology driven knowledge discovery process

A integração entre a engenharia de ontologias e KDD é uma tendência recente (ROTONDO; QUILLIGAN, 2020). Através de ontologias, esta integração é capaz de gerar abstrações amplas dos processos de descoberta de conhecimento. Ontologias de domínio podem ser utilizadas para melhorar a compreensão de um problema e suportar análises do tipo *hypothesis-driven*. Esta extensão do método KDD também visa a aplicação de métodos de mineração de dados a serem realizadas de forma automática ou semi-automática para obtenção de conhecimento à partir dos dados (GOTTGTROY, 2007).

Tradicionalmente, ontologia é um ramo da metafísica que se ocupa em definir quais categorias de entidades existem e suas relações. Em Ciências da Computação, o conceito de ontologia é aplicado em processos de modelagem, seja em banco de dados ou representação do conhecimento (ALMEIDA, 2014). Desta forma, ontologias tratam da representação, nomeação e da definição de categorias, e das propriedades e relações entre conceitos, dados e outras entidades.

Como resume Almeida (2014), ontologia em Ciência da Informação pode ser vista como um sistema conceitual informal, realizado através da "criação de vocabulários controlados para recuperação da informação a partir de documentos".

Conforme afirma Gottgtroy (2007), a integração de ontologias e ao método KDD permite o acúmulo gradativo de conhecimento sobre bases de dados, de forma a se obter um conhecimento de campo amplo através da aplicação iterativa do KDD.

Neste sentido, Gottgtroy (2007) propõe o *Ontology Driven Knowledge Discovery* (ODKD), como uma metodologia e modelo de processos que define, em diferentes níveis, as relações entre ontologias os processos de KDD. Para tanto, o ODKD utiliza de tarefas emprestadas da metodologia CRISP-DM (a ser tratada a seguir, na p. 56) e as adapta aos conceitos de ontologias.

3.2 SEMMA

O modelo de processos SEMMA, desenvolvido pelo Instituto SAS, representa o acrônimo *Sample, Explore, Modify, Model, Assess*, sendo aplicado na condução de projetos de mineração de dados, sendo introduzido em 1997. Detalhando seus passos (adaptado de Azevedo e Santos (2008)):

***Sample* (Amostragem):** Consiste em um processo de amostragem de dados que permita obter um subconjunto significativo à partir de um grande conjunto de dados, em um tamanho de amostra pequeno o suficiente que permita a sua rápida manipulação.

***Explore* (Exploração):** Análise exploratória da amostra de dados, em busca de tendências, anomalias e padrões, provendo um conhecimento gradativo sobre a base de dados.

***Modify* (Modificação):** Modificação do conjunto de dados através da seleção de variáveis e criação de novas variáveis (enriquecimento da base de dados) visando a etapa de modelagem.

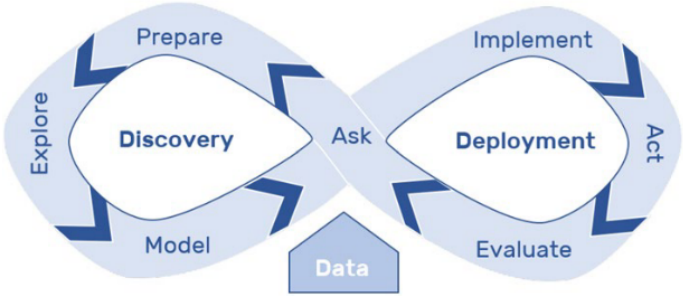
***Model* (Modelagem):** Realização da modelagem de dados pelo software/pacote estatístico, de modo a selecionar automaticamente um conjunto de variáveis que consigam predizer com qualidade um desfecho.

***Assess* (Interpretação):** Interpretação dos resultados do modelo, visando estabelecer sua utilidade e confiabilidade.

O SEMMA pode ser visto como uma implementação do modelo de processos KDD, existindo um paralelo entre os modelos. Ainda que possa ser aplicado de forma iterativa, não há uma definição explícita disto em sua definição.

Em 2016, o modelo foi revisto apresentando uma aplicação explicitamente cíclica (SAS, 2016), conforme evidencia a figura 5.

Figura 5 – Modelo de processos SEMMA revisto (2016)



Fonte: Rotondo e Quilligan (2020)

O modelo passa a apresentar os seguintes passos (adaptado de Rotondo e Quilligan (2020)):

***Ask a question* (Faça uma pergunta):** Compreensão do negócio, definição de objetivos e questões a serem respondidas.

***Prepare the data* (Prepare os dados):** Reúna e integre dados de diferentes fontes, criando um conjunto de dados válido para modelos analíticos.

***Explore the data* (Explore os dados):** Análise exploratória dos dados, utilizando ferramentas de visualização de dados, visando o refinamento das questões e objetivos.

***Model de data* (Modelagem dos dados):** Aplicação de vários algoritmos de modelagem e de aprendizado de máquina visando responder as questões do negócio.

***Implement your models* (Implemente os modelos):** O conhecimento ganho com a modelagem é posto em prática, redefinindo processos que podem ser automatizados.

***Act on new information* (Ação com novas informações):** Estando o modelo aprovado, decisões estratégicas e operacionais podem ser criadas ou revistas.

***Evaluate your results* (Avalie seus resultados):** Com base nos resultados advindos das modificações, avalie-os frente aos objetivos do negócio.

***Ask again* (Pergunte novamente):** Este passo encerra o ciclo, convidando a avaliar novamente a compreensão do negócio, objetivos e questões.

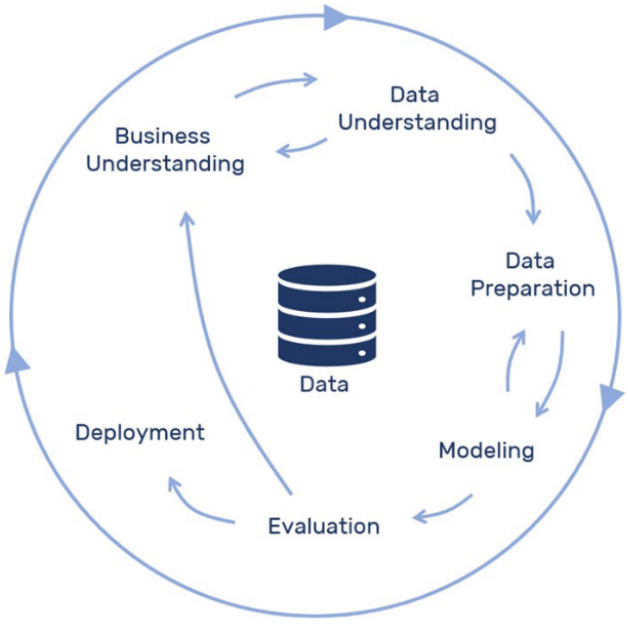
O modelo de processos SEMMA revisto ainda se baseia nos processos originais do KDD, enfatizando aspectos cíclicos e da aplicação do conhecimento obtido no negócio (*deployment*)

3.3 CRISP-DM

O modelo de processos CRISP-DM foi desenvolvido por um consórcio de empresas (SPSS, NCR, DaimlerChrysler e OHRA) e representa o acrônimo *Cross-Industry Standard Process for Data Mining*. As duas primeiras empresas são do ramo de pacotes estatísticos e banco de dados, equanto as duas últimas forneciam dados e casos de estudo. Sua primeira versão foi anunciada em 2000 e pode ser representado pela figura 6. CRISP-DM é um dos modelos mais bem documentados, e talvez por este motivo, um dos mais amplamente utilizados (ROTONDO; QUILLIGAN, 2020).

Ele é composto por 6 estágios (adaptado de Azevedo e Santos (2008)):

Figura 6 – Modelo de processos CRISP-DM



Fonte: Rotondo e Quilligan (2020)

***Business understanding* (Compreenção do negócio):** Compreensão dos objetivos do projeto em uma perspectiva de negócios, convertendo este conhecimento em definições de problemas de mineração de dados e em um plano inicial para resolução destes problemas.

***Data understanding* (Compreensão dos dados:** Acesso inicial aos dados, de modo a se familiarizar com os conjuntos de dados disponíveis, identificar possíveis problemas de qualidade nos dados, primeiras impressões e *insights* para a formulação de hipóteses a serem verificadas.

***Data preparation* (Preparação dos dados):** Atividades de modificação dos dados para prepará-los para a etapa seguinte de modelagem.

***Modeling* (Modelagem):** Aplicação de várias técnicas de modelagem e calibração de seus parâmetros.

***Evaluation* (Avaliação):** Interpretação dos resultados dos modelo, suas qualidades e limitações.

***Deployment* (Implementação):** À partir dos modelos construídos, implementar no negócio original o conhecimento obtido, de modo a dar conta dos problemas inicialmente identificados.

O modelo CRISP-DM utiliza os elementos mais importantes do modelo KDD e os insere em um processo cíclico que prevê explicitamente a compreensão do negócio e a implementação de alterações no negócio segundo os resultados obtidos (*deployment*).

3.4 Pontos em comum entre os modelos de processos

Os modelos de processos KDD, SEMMA e CRISP-DM possuem alguns pontos em comum. Em geral, são evoluções do mesmo esforço de formalizar os passos necessários para a aplicação de mineração de dados, prevendo passos anteriores e posteriores a efetiva modelagem dos dados.

Azevedo e Santos (2008) pontua as correspondências entre estes três modelos de processos. Considerando a atualização do modelo de processos SEMMA (ROTONDO; QUILLIGAN, 2020), apresenta-se o quadro 3.

Pode-se observar que o modelo SEMMA revisto apresenta um maior número de passos detalhados, enquanto que o CRISP-DM apresenta uma condensação destes passos em diretivas mais abrangentes.

Tanto o SEMMA revisto quanto o CRISP-DM apresentam etapas anteriores e posteriores aos passos do modelo KDD, voltadas à compreensão do problema e retorno ao passo inicial, formalizando assim ambos os processos como ciclo constantes.

Quadro 3 – Correspondências entre os modelos de processo KDD, SEMMA e CRISP-DM

KDD	**SEMMA revisto**	**CRISP-DM**
Pré-KDD	Faça uma pergunta	Compreensão do negócio
Seleção	Prepare os dados	Compreensão dos dados
Pré-processamento	Explore os dados	
Transformação		Preparação dos dados
Mineração de dados	Modelagem dos dados	Modelagem
Interpretação/Avaliação	Implemente os modelos	Implementação
	Ação com novas informações	
	Avalie seus resultados	Avaliação
Pós-KDD	Pergunte novamente	

Fonte: adaptado de Azevedo e Santos (2008) e Rotondo e Quilligan (2020).

3.5 KDD-PH: um modelo de processos para mineração de dados em Saúde Pública

O desenvolvimento dos modelos de processo KDD, SEMMA e CRISP-DM se formaram tomando o ambiente corporativo e de negócios como alvo. Desta maneira, termos como "cliente", "negócio" e "*deploy*" são comuns em suas descrições.

Considerando a aplicação destes processos da *ciência de dados* em Saúde Pública, alguns destes preceitos podem ser adaptados e convertidos, formando assim um modelo de processos dedicado à mineração de dados em saúde.

Abaixo são sugeridos passos de um modelo de processos para mineração de dados em Saúde Pública considerando o contexto acadêmico de pesquisa como contexto, denominado *Knowledge Discovery in Databases for Public Health* (KDD-PH).

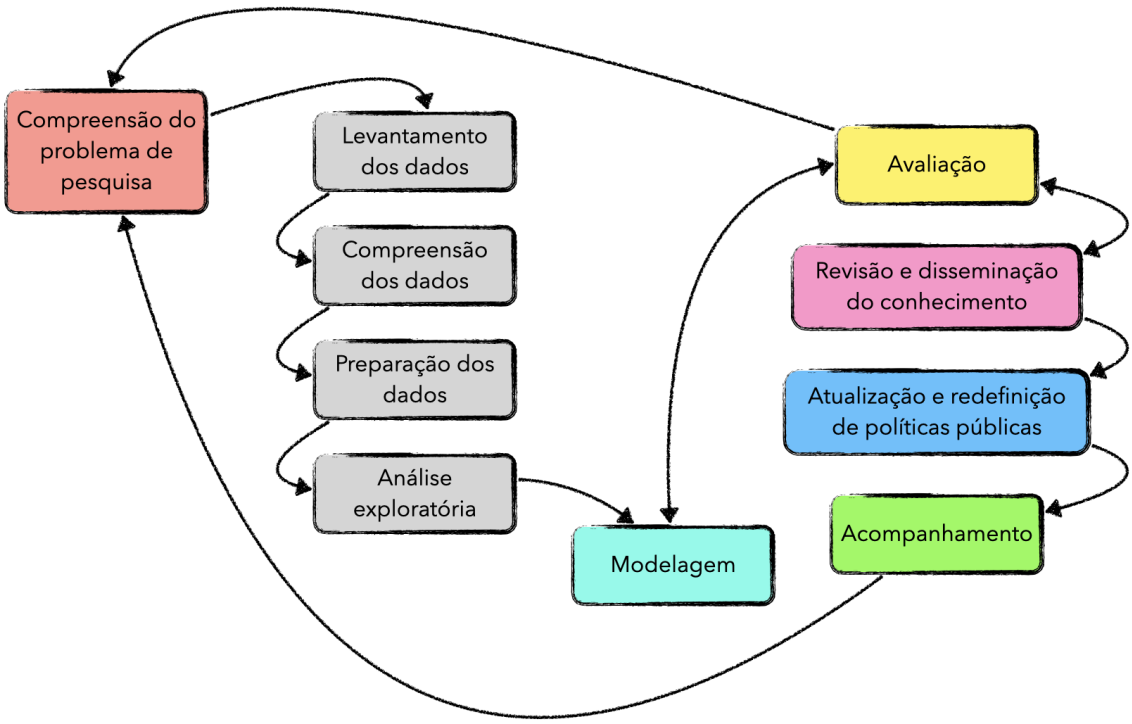
Compreensão do problema de pesquisa Definição e compreensão do problema de pesquisa em saúde. Levantamento bibliográfico sobre o tema e definição do processo saúde-doença.

Levantamento dos dados Levantamento das bases de dados secundárias disponíveis e acessíveis que podem auxiliar a modelar o problema de pesquisa. Pedidos de acesso à bases de dados restritas de interesse. Levantamento em campo através de pesquisas e inquéritos para obtenção de dados primários.

Compreensão dos dados Acesso inicial aos dados e seus dicionários, avaliação da qualidade e possibilidades de interligação entre as bases de dados. Formulação das primeiras hipóteses a serem verificadas.

Preparação dos dados Conversão dos dados em formatos computacionais comuns, possibilidades de armazenamento dos dados em banco de dados relacionais (SQL) ou índices de pesquisa (NoSQL). . Rotulagem de variáveis categóricas, re-codificação de

Figura 7 – Modelo de processos KDD-PH



Fonte: elaborado pelo autor.

variáveis, seleção de variáveis, criação de novas variáveis e enriquecimento dos dados. *Linkage* entre bases de dados.

Análise exploratória Análise exploratória dos dados preparados, verificação de possíveis inconsistências, criação e redefinição de hipóteses.

Modelagem Verificação das hipóteses, detecção de padrões e comportamentos. Mineração de dados.

Avaliação Interpretação dos resultados do modelo, suas qualidades, limitações e capacidade de responder o problema de pesquisa estabelecido. Possibilidade de retorno às etapas iniciais, remodelando-se o problema de pesquisa, levantamento de dados ou preparação dos dados.

Revisão e disseminação do conhecimento Publicação do conhecimento gerado em periódicos científicos para revisão por pares dos conhecimento obtidos, possíveis falhas e melhorias. Disseminação do novo conhecimento em canais externos à academia.

Atualização e redefinição de Políticas de Saúde Atuação e redefinição das políticas públicas frente ao novo conhecimento gerado.

Monitoramento e avaliação Acompanhamento do problema de pesquisa, novos conhecimentos gerados por outros grupos de pesquisa, atuação das políticas de saúde e seus resultados práticos.

Compreensão do problema de pesquisa

Etapa comum em projetos de pesquisa acadêmica, o problema de pesquisa deve ser delimitado à partir de um tema, sendo realizado um levantamento bibliográfico de pesquisas que já potencialmente abordaram este mesmo problemas. Com base no levantamento bibliográfico existente, o processo saúde-doença em questão deve ser estudado provisoriamente, com a previsão de seus espaços de determinantes e condicionantes. (CASTELLANOS, 1990).

O entendimento do processo saúde-doença é provisório pois ainda passará por etapas seguintes de verificação e estudo, até que esteja devidamente estabelecido e fundamentado.

Levantamento dos dados

Esses conhecimentos prévios sobre o processo saúde-doença permitem identificar fontes de dados e variáveis que podem potencialmente conter informações relevantes sobre este processo devem ser levantadas.

Recomenda-se iniciar este levantamento à partir dos Sistemas de Informação em Saúde já estabelecidos, conforme abordado na subseção Coleta e seleção de dados (p. 43). Fontes de dados não tradicionais também devem ser levado em consideração, como bases de dados oriundas de redes sociais.

Nesta etapa, deve-se também prever o trâmite necessário para a requisição e acesso a bases de dados restritas, seja através dos canais existentes nas instituições para tal fim ou através do uso ciadão da Lei de Acesso à Informação.

O levantamento primário de dados, através de pesquisas e inquéritos não deve ser ignorado, posto que para responder certas questões de pesquisa, pode ser necessário o levantamento direto destas respostas.

Compreensão dos dados

Após o levantamento dos dados que possam estar relacionados ao processo saúde-doença em questão, um acesso inicial a estes dados deve ser realizado, tão como ao seu dicionário de dados e documentações necessárias para sua posterior creditação. A verificação da qualidade dos dados acessados, como por critérios de completude, duplicidade e consistência devem ser considerados, como exemplicicado na subseção Pré-processamento (p. 45).

Preparação dos dados

Os dados de saúde pública costumam ser fornecidos em uma grade variedade de formatos. Como sugerido na subseção Transformação (p. 48), pode ser produtiva a conversão e convergência dos dados obtidos para um formato e ambiente comum de trabalho.

Etapas de rotulagem dos dados categóricos, com base no dicionário de dados fornecido, devem ser realizadas nesta etapa. Em seguida, a base de dados deve ser dimensionada para a investigação do processo saúde-doença em questão, selecionando as variáveis pertinentes e criando outras variáveis se possível, enriquecendo o banco de dados base com informações secundárias de possível importância.

O junção de bases de dados (*linkage*), de maneira probabilística ou direta, pode também ser prevista nesta etapa, visando ao fim a constituição de uma base de dados adequada à investigação do processo saúde-doença.

Análise exploratória

Após a formação de uma base de dados, a análise exploratória dos dados disponíveis pode ser realizada, buscando-se possíveis inconsistências (que podem ser advindas seja dos dados originais ou dos processos de preparação dos dados). Através de gráficos e ferramentas de visualização, as hipóteses podem ser redefinidas à luz dos dados existentes e disponíveis.

Modelagem

Nesta etapa, as hipóteses formuladas e refinadas podem ser verificadas, tão como novas hipóteses podem ser levantadas através dos métodos de *data mining* aplicados, conforme a subseção Mineração de dados (p. 49).

Avaliação

Os resultados obtidos na etapa anterior devem ser avaliados, buscando-se medir a qualidade do modelo, suas limitações e sua capacidade de modelar o processo saúde-doença em questão. Métodos de visualização podem ser aplicados nesta etapa, como previsto na subseção Visualização (p. 50).

Neste passo, pode-se cogitar objetivamente o retorno aos passos iniciais caso os resultados obtidos não estejam alinhados com os objetivos de pesquisa elencados. Os objetivos podem ser repensados ou os métodos adotados podem ser redefinidos.

Revisão e disseminação do conhecimento

Com os modelos calibrados e resultados avaliados e comentados, segue uma etapa comum à pesquisa acadêmica: a revisão e disseminação do conhecimento obtido na pesquisa.

A revisão do conhecimento se dá através da avaliação por pares de todo o processo envolvido. Nesta etapa, possíveis vieses da pesquisa podem ser apontados, novos dados podem ser sugeridos, alterações no método de modelagem podem ser recomendadas e interpretações falhas dos resultados podem ser reconduzidas.

Após a publicação científica do conhecimento gerado, este deve ser disseminado para fora da acadêmia, visando a população em geral, os gestores da saúde e demais atores envolvidos no processo saúde-doença em questão. Para tanto, a linguagem e formato de divulgação deve ser adaptada ao público visado.

Atualização e redefinição de Políticas de Saúde

Com o novo conhecimento sobre o processo saúde-doença estabelecido cientificamente e devidamente divulgado, as Políticas de Saúde pertinentes podem ser atualizadas e redefinidas.

Acompanhamento

A Ciência produzida por uma pesquisa deve, necessariamente, procurar produzir frutos e impactos na sociedade, não devendo ter um fim em si mesma. Desta forma, o acompanhamento das consequências das do novo conhecimento deve ser realizado, visando saber como este conhecimento está sendo utilizado na prática e possíveis possibilidades de atualização deste conhecimento frente a novas pesquisas realizadas.

A seguir, no capítulo Resultados, são apresentados artigos científicos e um capítulo de livro que utilizaram passos semelhantes ao KDD, SEMMA e CRISP-DM, tendo contribuído para à construção do KDD-PH.

4 Resultados

Apresentam-se neste capítulo a produção bibliográfica produzida para compor esta tese. Como pontos em comum, estas publicações advém do mesmo Referencial teórico e utilizam modelos de processos de análise de dados apresentados na Metodologia, tal como são contribuições da *ciência de dados em saúde* para a *Saúde Pública* e foram ou serão publicadas em formato aberto. A produção bibliográfica apresentada como Resultados desta tese foi dividida em cinco sessões, a seguir.

São apresentados cinco artigos, dos quais três foram publicados, um foi aprovado para publicação e um será submetido. Também é apresentado um capítulo de livro que se encontra em fase de editoração, totalizando assim seis produções bibliográficas.

As produções bibliográficas são reproduzidas em sua íntegra, mantendo a numeração de páginas original. Para diferenciação entre produção bibliográfica e texto original desta tese, as páginas que reproduzem as produções bibliográficas são envoltas por margens na cor preta.

4.1 Construção teórica

Como uma produção exclusivamente teórica sobre *ciência de dados* e *big data* na saúde, apresenta-se um artigo publicado na Cadernos de Saúde Coletiva, entitulado "Ciência de dados e bigdata: o que isto significa para estudos populacionais e da saúde". O artigo foi elaborado para uma edição especial do periódico sobre contribuições de *ciência de dados* e *big data* para os campos da saúde e demografia.

O artigo percorre os fundamentos de *ciência de dados* e *big data*, procurando destacar suas possíveis contribuições para os campos da saúde e demografia, com proposições sobre possibilidades de uso destes métodos e uma avaliação breve das contribuições nacionais existentes.

O artigo é reproduzido abaixo, em formatação simples, contendo o mesmo texto que foi aprovado pelos pareceristas em 8 de janeiro de 2020. Em seguida, após o artigo, é reproduzida a carta de aceite dos editores para publicação.

Ciência de dados e bigdata: o que isto significa para estudos populacionais e da saúde

Raphael Saldanha, Christovam Barcellos, Marcel Pedroso

Aceito em 8 de janeiro de 2020,
na Cadernos de Saúde Coletiva

Resumo

Introdução. O termo *big data* no ambiente acadêmico tem deixado de ser uma novidade, tornando-se mais comum em publicações científicas e em editais de fomento à pesquisa, levando a uma revisão profunda da ciência que se faz e se ensina. O objetivo deste artigo é refletir sobre as possíveis mudanças que as ciências de dados podem provocar nas áreas de estudos populacionais e de saúde. *Métodos.* Para fomentar esta reflexão, artigos científicos selecionados da área de *big data* em saúde e demografia foram contrastados com livros e outras produções científicas. *Resultados.* Argumenta-se que o volume dos dados não é a característica mais promissora de *big data* para estudos populacionais e saúde, mas a complexidade dos dados e a possibilidade de integração com estudos convencionais através de equipes interdisciplinares é promissora. *Conclusão.* No âmbito do setor saúde e de estudos populacionais, as possibilidades da integração dos novos métodos de ciência de dados aos métodos tradicionais de pesquisa são amplas, incluindo um novo ferramental para análise, monitoramento, predição de eventos (casos) e análises processos de saúde-doença na população e estudo dos determinantes socioambientais e demográficos.

Palavras-chave: saúde pública; demografia; ciência de dados; *big data*.

1 Introdução

O termo *big data* no ambiente acadêmico tem deixado de ser uma novidade, tornando-se mais comum em publicações científicas[1] e em editais de fomento à pesquisa[2]. Departamentos de universidades e centros de pesquisa internacionais e nacionais têm redesenhado suas ementas ou criado programas e disciplinas para atender a demanda de formação em ciência de dados e *big data*[3]. Passado o *hype* da novidade, resta refletir sobre as mudanças necessárias e seus impactos nas áreas de estudos populacionais e de saúde.

Os sistemas de informação de saúde e os recenseamentos, inquéritos e outras pesquisas demográficas são as principais fontes de dados para o conhecimento das dinâmicas populacionais. Tais pesquisas já produzem um respeitoso volume de dados, introduzindo complexidades para a sua análise. Já éramos *big data*? Ou *big data* se define não só pela quantidade de dados em análise?

Diversas propostas para definição do que vem a ser *big data* (ou sua vertente científica e acadêmica denominada "ciência de dados") se apresentam, acompanhando naturalmente a evolução de um campo novo, absorvendo

[1]Levantamento não sistemático realizado pelos autores em junho de 2019 no *Web of Science* via no Portal de Periódicos da CAPES utilizando os descritores nos campos "título" e "assunto" com as palavras "*big data*" e "*health*" e com filtro de tópico "*big data*" ativado retornou 3.728 revisados por pares em revistas científicas indexadas, sobretudo a partir de 2012.

[2]Iniciativa recente foi a chamada conjunta do *Grand Challenges Explorations* voltada exclusivamente e pela primeira vez a pesquisadores brasileiros, é resultado da parceria entre o Ministério da Saúde (MS), o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), o Conselho Nacional das Fundações Estaduais de Amparo à Pesquisa (CONFAP), as Fundações Estaduais de Amparo à Pesquisa (FAPs) e a Fundação Bill & Melinda Gates (FBMG). Ver detalhes em ¡https://gcgh.grandchallenges.org/challenge/grand-challenges-explorations-brazil-data-science-approaches-improve-maternal-and-child¿.

[3]Alguns exemplos internacionais e nacionais: Harvard Data Science Initiative; LSE: Data Science; University of California: MIT: Data Science and Big Data Analytics Making Data-Driven Decisions; Master of Information and Data Science; University of Denver Master of Science in Data Science; Syracuse University: Master in Data Science; Columbia University: Master of Science in Data Science; Universidade de Lisboa: Ciência de Dados; USP: São Paulo School of Advanced Science on Learning from Data; UNICAMP: Difusão em Ciência de Dados; PUC RIO: Ciência de Dados; PUC Minas: Ciência de Dados e Big Data; CEFET/RJ: Ciência da Computação com ênfase em Ciência de Dados; Icict/Fiocruz: Ciência de Dados aplicada à Saúde; Farmanguinhos/Fiocruz: Big Data em Saúde; LNCC: Programa Multidisciplinar de Pós-Graduação do Laboratório Nacional de Computação Científica.

as suas rápidas mudanças e novas aplicações. Contudo, parece haver uma convergência: *big data* não diz respeito, unicamente, a grandes bases de dados. As implicações da correção desta característica para estudos populacionais e de saúde são importantes. O volume e velocidade da produção de dados destas áreas em comparação aos campos das ciências naturais, como física e biologia é menor, mesmo comparando-se todos os microdados de todas as pesquisas demográficas e dos sistemas de informação em saúde.

Nesse contexto, o termo "ciência de dados" vem se consolidando como um campo de convergência tecnológica, científica e acadêmica, filosófica e pragmaticamente interdisciplinar, formado basicamente por cientistas da computação, matemáticos, estatísticos e pesquisadores com o conhecimento substantivo do problema em análise - como os médicos e sanitaristas no caso na saúde, mas podemos incluir aqui grupos de pesquisas em ciência de dados que contam com biólogos, geneticistas, economistas, financistas, geógrafos, advogados, historiadores, entre outros profissionais em suas equipes).

A definição que utilizamos em nosso grupo de pesquisa[4] para esse domínio em construção é: "Ciência de Dados é um campo de estudo que se destaca pela capacidade de auxiliar a descoberta de informação útil a partir de grandes ou complexas bases de dados, bem como a tomada de decisão orientada por dados" [13].

Desta forma, entendemos que *big data* é um dos aspectos do campo da ciência de dados (que trata de outros aspectos como estratégias para extração, transformação e carga dos dados, modelagem, construção e avaliação de algoritmos descritivos e preditivos, visualização de grandes quantidades de dados e *deploy* dos modelos em ambientes de produção para tomada de decisão, entre outros) e o que importa na definição de *big data* não é o volume ou mesmo velocidade da produção de dados, mas a complexidade estrutural desses dados (variedade) e o poder computacional necessário para analisá-los integralmente.

Esta complexidade deriva, não só da multiplicidade de fontes de dados e de variáveis inter-relacionadas, mas também da estrutura sociodemográfica contemporânea, imbricada por fatores econômicos, culturais e políticos. A necessidade de se atribuir grandes categorias sociais às populações tem, de fato, dominado o debate tanto das ciências sociais, quanto o discurso político, que costuma utilizar termos simplificadores para designar a classe pobre (des-

[4]Grupo de pesquisa em Ciência de Dados aplicada à Saúde, cadastrado no CNPq e certificado pela Fiocruz: `http://dgp.cnpq.br/dgp/espelhogrupo/4230691756969719`

camisados, excluídos, vulneráveis, despossuídos, miseráveis, *sans culottes*, proletários, entre outros), a classe média (pequena burguesia, emergentes, intelectuais, servidores públicos...) e a mais rica também (elite, patronado, classe alta, etc.).

Na sociologia, também os recortes teóricos são produzidos segundo concepções ideológicas, com marcadas contribuições de escritos clássicos de Marx, Durkheim e Weber [9]. Se para Marx, a estrutura social capitalista é resultado das relações de produção, que gera uma sociedade de classes antagônicas e complementares, para Weber, a estratificação social deriva da distribuição desigual de riqueza, prestígio e poder. Mais recentemente, Bourdieu introduziu o conceito de capital cultural, identificando distinções de classe pelos seus valores, hábitos e padrões de consumo de bens e serviços [4], deslocando o debate sobre inclusão/exclusão social da esfera da produção para a esfera da circulação e consumo.

Enquanto alguns destes conceitos podem ser operacionalizados para a coleta e análise de dados sociodemográficos e traduzidos em perguntas objetivas que irão compor os questionários de censos, inquéritos e os registros de saúde, como a posse de bens, cor e raça, sexos (mas não gênero), nível educacional e ocupação, outros são de difícil apreensão, como a detenção de meios de produção, identidade de classe e etnicidade, o prestígio social e a inserção nos circuitos de poder, econômicos e culturais.

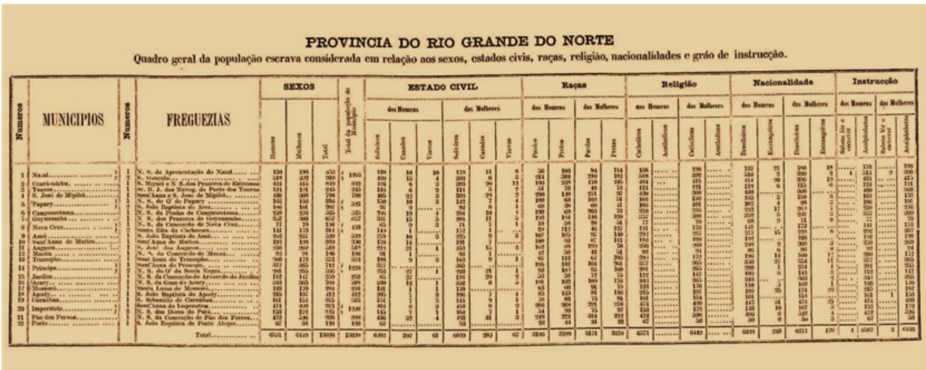
Qualquer que seja a abordagem teórica sobre a estrutura social, é evidente que esta estrutura e estratificação constitui um fenômeno emergente da sociedade de classes e a posição social de cada indivíduo é definida pela sua inserção política, econômica e simbólica nesta sociedade. No caso dos estudos quantitativos, estes componentes somente podem ser identificados a posteriori, isto é, pela análise integrada e multidimensional de uma variedade e quantidade de dados que captam os diversos aspectos desta inserção do indivíduo na sociedade [11]. Este tipo de análise pode ser viabilizado pelas técnicas estatísticas, pelo manuseio criativo das fontes de dados tradicionais, pela incorporação de novas fontes de dados e pela infraestrutura computacional atualmente disponível.

Não por acaso, os primeiros censos demográficos geraram tabelas em geral bivariadas, contendo a quantidade de habitantes segundo sexo, estado civil, nacionalidade, religião, grau de instrução e profissão, com uma marcada diferença entre pessoas "escravas e livres". Novos censos, realizados principalmente após 1970, introduziram diversas outras variáveis e instrumentos

de coleta de dados[5]. A partir destes novos dados foi possível criar outras formas de agregação de dados e realizar análises estatística multivariadas.

Mas a complexidade da sociedade brasileira, entre outros fatores, exigem novas estratégias de categorização e análise. As análises estatísticas convencionais tendem a identificar correlações ou associações entre variáveis e grandes padrões de semelhança de grupos populacionais. As técnicas de *data mining* e a diversidade de variáveis atualmente disponíveis permitem detectar disjunções, isto é, combinações entre variáveis que conformam grupos minoritários, militantes, projetos de vida alternativos e desconexões, que fogem aos padrões sociais hegemônicos e podem ser capturados mediante a "mineração" de fontes alternativas de informação como as redes sociais e outras formas de interação entre usuários via aplicativos.

Figura 1: Dados da "Provincia do Rio Grande do Norte"


PROVINCIA DO RIO GRANDE DO NORTE

Quadro geral da população escrava considerada em relação aos sexos, estados civis, raças, religião, nacionalidades e grão de instrucção.

Fonte: IBGE, `https://memoria.ibge.gov.br/sinteses-historicas/historicos-dos-censos/censos-demograficos.html`

Estão assentadas todas as condições tecnológicas para se trabalhar com o par de conceitos dialeticamente ligados de desigualdade e diferença, ou inclusão e exclusão [6]. Pois são estas categorias analíticas que podem explicar a vulnerabilização e marginalização de grupos sociais, bem como elucidar os diferentes graus de exposição aos determinantes socioambientais do processo saúde-doença intragrupos populacionais.

[5] `https://memoria.ibge.gov.br/sinteses-historicas/historicos-dos-censos/censos-demograficos.html`

2 A busca de padrões e divergências

Nas ciências da natureza, mais dados significam basicamente, maiores computadores ou necessidade de métodos que consigam lidar eficientemente com mais dados. Até certo momento, as ciências sociais também seguiam essa lógica.

Tradicionalmente, após a definição da pergunta de estudo e da área de estudo, se fazia um recorte das bases de dados, levando-se apenas as variáveis de interesse já previamente sugeridas por levantamentos bibliográficos ("pela literatura"). Estas variáveis são confrontadas em um modelo estatístico de verificação de hipóteses a priori que busca responder à pergunta inicial do estudo (*hypothesis-driven approaches*).

A inovação metodológica promissora na utilização de técnicas de *big data* nas ciências sociais e particularmente na saúde pública é permitir se fazer uma ciência a posteriori. Procurar padrões, localizar associações, visualizar a complexidade dos fenômenos, prever desfechos em saúde sem considerar previamente hipóteses formuladas a priori, predizer comportamentos com precisão sem necessariamente partir de uma sustentação teórica ou clínica. Transitar de "*theory-driven approach*" ou "*hypothesis-driven approach*" para "*data-driven approach*". Conforme afirma [3]:

> "*Our problems also require new ways of thinking about the data we have and new methods for analysing and visualizing these data. [...] We cannot simply scale up; we have to change almost everything*" [3].

Trabalhar com *big data* em estudos populacionais e saúde pública não significa necessariamente trabalhar com muitos dados, mas diz respeito a alterar profundamente o modo de se fazer pesquisa, o que leva a necessidade de reformulação dos currículos de graduação e pós-graduação. Se faz necessário ir além da mudança de nomes de disciplinas e departamentos, para uma revisão profunda da ciência que se faz e se ensina.

Aos críticos e céticos quanto a viabilidade da adoção disruptiva da abordagem *data-driven* em estudos populacionais e de saúde, cabe salientar que a discussão epistemológica não pode ser baseada em visões dicotômicas da ciência do tipo "qualitativo ou quantitativo" encaradas, frequentemente, como abordagens antagônicas. A proposta aqui defendida é de construir

caminhos para uma ciência aberta[6], criativa, inovadora, que possa adotar simultaneamente métodos mistos (quali e quanti) e que possam ser guiadas por procedimentos híbridos (hypothesis e data-driven) [8, 16, 7, 17].

Há uma evidente associação entre as variáveis socioeconômicas coletadas pelos sistemas de informação de saúde. A disponibilização de dados e o interesse retomado da epidemiologia pela busca da equidade em saúde fez crescer o número de estudos sobre desigualdades sociais e saúde [2]. Renda, educação, etnicidade e ocupação se manifestam na sociedade brasileira como uma conjunção de fatores que podem acarretar em melhores ou piores condições de vida e de saúde.

A adoção de grandes categorias e o uso de indicadores compostos e sintéticos têm sido empregados para a detecção destas desigualdades [2]. No entanto, estas abordagens não são capazes de responder ao efeito sobre a saúde de contextos particulares de risco, como grupos indígenas com crenças, hábitos e atitudes tradicionais que produzem perfis epidemiológicos característicos, e os afastam dos padrões urbanos e das categorias classicamente usadas para estudos sobre desigualdades. O que dizer, por exemplo sobre os idosos urbanos, que podem possuir uma renda minimamente necessária para seu sustento, mas encontram-se isolados devido a suas condições clínicas e familiares? Teriam eles um perfil diferenciado de risco? Mais que isso. Como vivem as mulheres pobres e negras da periferia das grandes cidades, submetidas a uma sobreposição de condições adversas de violência, dificuldades de acesso a bens e serviços e sujeitas a discriminações na cidade e nas instituições? Apontar os perfis particulares de risco de cada grupo populacional pode ser um importante passo para a construção de políticas de saúde inclusivas e mais adequadas a cada grupo.

Além disso, é importante ressaltar que pessoas com perfis sociodemográficos semelhantes, mas que moram e circulam em lugares diferentes, podem apresentar perfis epidemiológicos discrepantes. O território onde as pessoas moram define grande parte das condições de vida, da produção de doenças, de acesso aos serviços de saúde e da organização social local, o que permite estabelecer laços de solidariedade e compartilhar recursos para sua proteção [15]. A restituição dos dados coletados em censos ou registros de

[6]Destaque para a adesão crítica e estratégica da Fiocruz ao movimento global de "Ciência Aberta" que em linhas gerais, propõe tornar a pesquisa científica acessível para todos. Na prática, significa eliminar obstáculos artificiais, especialmente os editoriais, legais e econômicos, à livre circulação do conhecimento científico. Acesse o conjunto de iniciativas em `https://portal.fiocruz.br/ciencia-aberta`.

saúde ao seu lugar de origem, realizada atualmente por procedimentos de georreferenciamento, permite resgatar o contexto em que se produzem os riscos e a intensidade e frequência das exposições da população ao determinantes sócioambientais, mas igualmente onde se devam promover o estabelecimento de sistemas de proteção social. Neste sentido, as análises de dados por técnicas de *big data* não podem prescindir de informações geográficas que permitem complementar e contextualizar eventos de saúde-doença que se expressam sempre no nível individual, com dados sobre o ambiente, as condições socioeconômicas, a presença de instituições e redes de apoio, que são características subjacentes ao território.

3 *Big data* em saúde, acesso a dados e direito ao sigilo

A forma como lidamos com os dados também deve ser alterada. Tradicionalmente, pesquisas sobre mortalidade usam, quase que exclusivamente, dados do Sistema de Mortalidade; pesquisas sobre migração usam dados do Censo; pesquisas sobre orçamento e renda buscam dados de inquéritos específicos. Contudo, a complexidade dos fenômenos humanos, da sua saúde e comportamento, transita livremente por estas bases de dados e as transborda. Para procurar dar conta desta complexidade, o mínimo a fazer é procurar trabalhar com estes dados de forma integrada e interoperável.

Dados de mortalidade, de natalidade, notificação de doenças, internações hospitalares e atendimentos ambulatoriais, dentre outros, são captados por sistemas de informação específicos que registram eventos em saúde (ou seria em doença!), desde o nascimento até o óbito em diversos estabelecimentos de saúde ou fora destes, em diferentes ocasiões. Conectar estes eventos (e dados) permitiria traçar o percurso de pessoas dentro do sistema de saúde, além de definir a história clínica de cada pessoa e suas situações de saúde e de doença. Uma mesma pessoa pode ser portadora de HIV, sofrer violência e ser internada em uma unidade de emergência. É importante que os serviços de saúde sejam informados sobre esta combinação de fatores para poder prestar a atenção médica e psicossocial adequada e oportuna. Além disso, o conjunto de histórias individuais permite estimar a prevalência de determinadas doenças crônicas, sobrevida e identificar suas comorbidades.

Iniciativas importantes estão, por meio de métodos criativos, mistos,

híbridos e intensivos em computação científica, buscando estimar dados e indicadores de saúde-doença por meio de dados públicos individualizados, porém não identificados.

A revista Lancet em 2016, publicou um editorial intitulado "*GBD 2015: from big data to meaningful change*" em que apresentava aos leitores alguns dos principais achados do estudo, as estratégias metodológicas para a construção das estimativas e a publicação dos resultados do número especial "*Global Burden of Disease Study 2015*", uma parceria entre o *Institute for Health Metrics and Evaluation* (IHME) e a *Lancet* [10].

Em nossa opinião, a iniciativa *Global Burden of Disease* (GBD) é atualmente o esforço (humano e computacional) mais abrangente em estudos populacionais, epidemiológicos e saúde que utiliza métodos híbridos, mistos e *big data* aplicados a morbimortalidade das principais doenças, agravos e fatores de risco para a saúde-doença em níveis global, nacionais e regionais. Disponibilizam os resultados dos estudos, publicizam as metodologias, fontes de dados utilizados para a construção de suas principais métricas[7] e suas tendências desde 1990, bem como ferramentas para visualizações interativas desses resultados, permitindo, desta forma, que pesquisadores sejam estimulados a fazer comparações entre as populações e buscar compreender os desafios para a saúde pública decorrentes dessas tendências. No estudo de 2017, foram avaliadas 359 causas de morbimortalidade prematuras, incidência e prevalência de doenças e anos vividos com incapacidade para 195 países e territórios. Para ter acesso aos estudos e publicações derivadas dos GBDs 2017, 2016, 2015, 2013 e 2010, ver `https://www.thelancet.com/gbd`.

No Brasil, um acordo em 2015 entre o Ministério da Saúde, a Universidade Federal de Minas Gerais (UFMG) e o *Institute for Health Metrics and Evaluation* (IHME) da Universidade de Washington resultou no projeto "GBD Brasil". O projeto visa constituir uma rede de colaboradores, com participação de pesquisadores brasileiros e técnicos do Ministério da Saúde, com o objetivo de dar apoio metodológico e avaliar as estimativas do estudo GBD em nível subnacional, bem como compilar e analisar a carga de doença no país e nos estados brasileiros. Os primeiros resultados do estudo no Brasil, no âmbito desse projeto, foram publicados em suplemento da Revista Brasileira de Epidemiologia em 2017 e materializa os esforços para estimar e analisar a carga de doença no Brasil e nos estados. Acesse o suplemento em `https://www.scielosp.org/toc/rbepid/2017.v20suppl1/`.

[7] *Years of life lost* (YLLs), *years lived with disability* (YLDs), e a combinação dos dois últimos: *disability-adjusted life-years* (DALYs).

Em 2018, o Projeto GBD Brasil publicou o estudo "*Burden of disease in Brazil, 1990–2016: a systematic subnational analysis for the Global Burden of Disease Study 2016*", que consolida, estima e analisa mudanças políticas, econômicas e epidemiológicas que afetaram o processo saúde-doença no país. Os resultados do estudo apresentam estimativas do GBD 2016 para expectativa de vida ao nascer, esperança de vida saudável, morbimortalidade específicas por causas (333), perda de saúde devido à morte ou incapacidade e fatores de risco para o Brasil, seus 26 estados e o Distrito Federal, de 1990 a 2016 [1].

Criado em 2017, o Laboratório de *big data* e Análise Preditiva em Saúde (LABDAPS) da Faculdade de Saúde Pública da Universidade de São Paulo (FSP/USP) tem o objetivo de desenvolver pesquisas que auxiliem na melhoria da atenção à saúde no Brasil. Os pesquisadores do laboratório trabalham na aplicação e no desenvolvimento de métodos de inteligência artificial (*machine learning*) a problemas importantes da área da saúde, como a análise de impacto de políticas públicas de saúde, a melhoria da qualidade da informação de saúde e a predição da ocorrência de doenças e óbitos. Ver detalhes e publicações em `https://sites.google.com/view/labdaps`.

A Plataforma de Ciência de Dados aplicada à Saúde (PCDaS) do Laboratório de Informação em Saúde (LIS) do Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (ICICT) da Fundação Oswaldo Cruz (Fiocruz), criada em 2016 em parceria com o Laboratório Nacional de Computação Científica (LNCC) do Ministério da Ciência, Tecnologia, Inovações e Comunicações (MCTIC), é um projeto de pesquisa e desenvolvimento tecnológico que tem como objetivo principal disponibilizar serviços tecnológicos e computação científica para armazenamento, gestão, análise, visualização e disseminação de grandes quantidades de dados de saúde e seus determinantes socioambientais para pesquisadores, docentes e discentes de instituições de ensino e pesquisa, bem como gestores governamentais. Mais informações em: `https://bigdata.icict.fiocruz.br/` Por outro lado, a vinculação (*linkage*) de dados socioeconômicos com os diversos eventos de saúde individuais pode abrir a empresas privadas (seguros de saúde ou empregadores) ou golpistas a oportunidade de identificar pessoas com maior risco de adoecer e morrer, ameaçar, estigmatizar, excluir ou até chantagear estas pessoas. Mesmo o comportamento de uma pessoa nas mídias sociais pode ser capturado e tratado de modo a delinear perfis de risco e probabilidades de adoecimento por meio de algoritmos[8], que constituem informações

[8]Ver descrição interessante de aplicação de modelo de *machine learning* para detecção precoce de depressão em tweets. Disponível em `https://bit.ly/3axRJPG`.

de grande valor de mercado [5]. Em países com sólida tradição democrática, como a Suécia, se dispõe de uma grande massa de dados sociodemográficos e de saúde que são *linkados* por meio de um código de identificação pessoal, mantidos anonimizados em instituições de governo e disponibilizados para pesquisadores e planejadores de políticas públicas [12]. A preocupação com o uso indevido de dados individuais não é, portanto, necessariamente uma particularidade de governos autoritários, mas de sociedades modernas, nas quais técnicos capacitados podem acessar dados que se encontram nas nuvens ou em servidores públicos. Nesse contexto, a necessidade de *linkage* entre bases de dados torna-se imprescindível para a pesquisa e tomadores de decisão, e com ela, as preocupações sobre confidencialidade e segurança[9].

O conservadorismo das instituições e a noção de propriedade feudal sobre "seus" dados e resultados são questões que costumam se esconder sobre os princípios nobres e necessários da ética em pesquisa. Assimilar tecnologias de *big data* não significa abandonar estes princípios, mas confrontar a necessidade de completo acesso e uso dos dados às restrições que são da ordem técnica, e não de ordem patrimonialista ou política.

4 Desafios e perspectivas futuras

O uso de técnicas de *big data* em saúde pública apresenta condições para a superação de modelos simplistas de classificação de riscos e identificação de desigualdades. Novas categorias analíticas podem ser buscadas, por meio de algoritmos que ajudem a definir com maior precisão grupos e situações de risco e vulnerabilidade. A rápida transformação sociodemográfica do Brasil exige que se analise seu impacto sobre a saúde da população sob a perspectiva de grupos socioespaciais particulares [14], o que inclui o envelhecimento e urbanização da população, as novas formas de exclusão, que não são necessariamente determinadas pela renda, e a permanência de situações históricas de segregação e produção de condições adversas de vida, saúde e doença.

No âmbito do setor saúde e de estudos populacionais não é difícil imaginar as possibilidades da abordagem com técnicas de ciência de dados e *big*

[9]Destaque para a iniciativa do Centro de Integração de Dados e Conhecimentos para Saúde (Cidacs/Fiocruz/Bahia) que conduz estudos e pesquisas baseados em projetos interdisciplinares originados na vinculação de grandes volumes de dados para ampliar o entendimento dos determinantes e das políticas sociais e ambientais sobre a saúde da população, ver detalhes em `https://cidacs.bahia.fiocruz.br/`.

data para análise, monitoramento, predição de eventos (casos) e situações de saúde-doença na população, bem como a associação destes com seus determinantes socioambientais e demográficos.

Por outro lado, observa-se no Brasil um cenário revolto no que toca a formulação de políticas públicas e dados abertos. São agora comuns os ataques a pesquisas essenciais como o Censo Demográfico e DETER, tentativas de retrocesso na divulgação de dados, acesso à informação pública e captura de instituições públicas detentoras de dados estratégicos por interesses privados. Enquanto a ciência avança, com a incorporação de novos métodos e possibilidades, se faz necessário resguardar no campo cívico e político a preservação da autonomia e capacidade de inovação dos meios e modos de se fazer ciência.

Referências

[1] F Marinho; et al. «Burden of disease in Brazil, 1990–2016: a systematic subnational analysis for the Global Burden of Disease Study 2016». Em: *The Lancet* 392 (10149 2018).

[2] JB Silva; MBA Barros. «Epidemiologia e desigualdade: notas sobre a teoria e a história». Em: *Rev Panam Salud Publica* 12 (6 2002), pp. 375–383.

[3] SA Bohon. «Demography in the Big Data Revolution: Changing the Culture to Forge New Frontiers». Em: *Population Research and Policy Review* 37 (3 2018).

[4] Pierre Bourdieu. *Poder simbólico*. Bertrand, 1992.

[5] E Vayena; M Salathé; LC Madoff; JS Brownstein. «Ethical Challenges of Big Data in Public Health». Em: *PLoS Comput Biol* 11 (2 2015).

[6] NG Canclini. *Diferentes, desiguais e desconectados*. UFRJ, 2009.

[7] ADP Chiavegatto. «Uso de big data em saúde no Brasil: perspectivas para um futuro próximo». Em: *Epidemiol Serv Saúde* 24 (2015), pp. 325–32.

[8] J Creswell. *Projeto de Pesquisa: Métodos Qualitativo, Quantitativo e Misto*. 4ª ed. Bookman, 2010.

[9] A Giddens. *Capitalism and Modern Social Theory: An Analysis of the Writings of Marx, Durkheim and Max Weber*. Cambridge University Press, 2009.

[10] The Lancet. «GBD 2015: from big data to meaningful change». Em: *The Lancet* 388 (10053 2016).

[11] F Lebaron. *How Bourdieu "quantified" Bourdieu: The geometric modelling of data. Quantifying Theory*. Springer, 2009.

[12] P Cnudde; O Rolfson; S Nemes. «Linking Swedish health data registers to establish a research database and a shared decision-making tool in hip replacement». Em: *BMC Musculoskelet Disord* 17 (1 2016), p. 414.

[13] PCDaS. *Plataforma de Ciência de Dados aplicada à Saúde. Laboratório de Informação em Saúde (Lis). Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (Icict). Fundação Oswaldo Cruz (Fiocruz)*. URL: `https://bigdata.icict.fiocruz.br`.

[14] C Barcellos; PC Sabroza; P Peiter; LI Rojas. «Organização espacial, saúde e qualidade de vida: análise espacial e uso de indicadores na avaliação de situações de saúde». Em: *Inf Epidemiol SUS* 11 (2002), pp. 129–38.

[15] L Rojas. «Geografía y salud: temas y perspectivas en América Latina». Em: *Cad. Saúde Pública* 14 (4 1998).

[16] G Shmueli. «To Explain or To Predict?» Em: *Statistical Science* (2010).

[17] KC Elliot; KS Cheruvelil; GM Montgomery; PA Soranno. «Conceptions of Good Science in Our Data-Rich World». Em: *Bioscience* 66 (10 2016), pp. 880–889.

De: **Raphael Guimarães** onbehalfof@manuscriptcentral.com
Assunto: Cadernos Saúde Coletiva - Decision on Manuscript ID CADSC-2019-0305.R1
Data: 8 de janeiro de 2020 16:30
Para: raphael.saldanha@icict.fiocruz.br

08-Jan-2020

Dear Mr. Saldanha:

It is a pleasure to accept your manuscript entitled "Demog Ciência de dados e big data: o que isto significa para estudos populacionais e da saúde" in its current form for publication in the Cadernos Saúde Coletiva. The comments of the reviewer(s) who reviewed your manuscript are included at the foot of this letter.

Thank you for your fine contribution. On behalf of the Editors of the Cadernos Saúde Coletiva, we look forward to your continued contributions to the Journal.

Sincerely,
Dr. Raphael Guimarães
Associate Editor, Cadernos Saúde Coletiva
raphael.guimaraes@fiocruz.br, raphael@iesc.ufrj.br

Entire Scoresheet:
Reviewer: 1

Recommendation: Accept

Comments:
(There are no comments.)

Additional Questions:
Does the manuscript contain new and significant information to justify publication?: Yes

Does the Abstract (Summary) clearly and accurately describe the content of the article?: Yes

Is the problem significant and concisely stated?: Yes

Are the methods described comprehensively?: Not applicable

Are the interpretations and conclusions justified by the results?: Yes

Is adequate reference made to other work in the field?: Yes

Is the language acceptable?: Yes

Length of article is: Adequate

Number of tables is: Adequate

Number of figures is: Adequate

Please state any conflict(s) of interest that you have in relation to the review of this paper (state “none” if this is not applicable).: None

Rating:

Interest: 1. Excellent

Quality: 2. Good

Originality: 1. Excellent

Overall: 1. Excellent

Reviewer: 2

Recommendation: Accept

Comments:
O artigo traz à tona uma importante discussão para a área de demografia e saúde pública. As correções feitas deixaram o artigo com um melhor entendimento

deixaram o artigo com um melhor entendimento.

Additional Questions:
Does the manuscript contain new and significant information to justify publication?: Yes

Does the Abstract (Summary) clearly and accurately describe the content of the article?: Yes

Is the problem significant and concisely stated?: Yes

Are the methods described comprehensively?: Not applicable

Are the interpretations and conclusions justified by the results?: Not applicable

Is adequate reference made to other work in the field?: Yes

Is the language acceptable?: Yes

Length of article is: Adequate

Number of tables is: Adequate

Number of figures is: Adequate

Please state any conflict(s) of interest that you have in relation to the review of this paper (state “none” if this is not applicable).:

Rating:

Interest: 2. Good

Quality: 2. Good

Originality: 3. Average

Overall: 3. Average

4.2 Captura e seleção de dados

Aqui são apresentados dois artigos que, em comum, se dedicam especificamente a tratar da captura e seleção de dados.

O primeiro artigo, denominado "Microdatasus: pacote para o download e pré-processamento de microdados do Departamento de Informática do SUS (DataSUS)", foi aprovado para publicação em 17 de junho de 2019. O artigo apresenta um pacote de funções para o pacote estatístico R desenvolvido pelo autor.

O pacote apresentado no artigo visa o *download* e pré-processamento de microdados dos Sistemas de Informação em Saúde disponibilizados pelo DataSUS, se relacionando diretamente com as etapas de coleta e seleção de dados, tão como a etapa de transformação de dados.

Segundo informações do periódico, a página do artigo na Internet recebeu 2.274 acessos entre Setembro de 2019 e Novembro de 2020.

O segundo artigo, denominado "Dataset of hospital admissions authorizations in Brazil", está em fase final de revisão para submissão.

O artigo foi desenvolvido em conjunto com a equipe da Plataforma de Ciência de Dados aplicada à Saúde (PCDaS/ICICT/Fiocruz) e segue o formato de *data paper*. Neste formato, são apresentados os passos utilizados para a criação de uma base de dados, no caso, as Autorizações de Internação Hospitalar – AIH, advindas do Sistema de Informação Hospitalar (SIH) do DataSUS.

Neste artigo, são apresentados em detalhe etapas de leitura da base de dados, re-codificação de variáveis e alimentação de um índice de pesquisa do tipo *NoSQL*.

CSP CADERNOS DE SAÚDE PÚBLICA
REPORTS IN PUBLIC HEALTH

QUESTÕES METODOLÓGICAS
METHODOLOGICAL ISSUES

Microdatasus: pacote para *download* e pré-processamento de microdados do Departamento de Informática do SUS (DATASUS)

Microdatasus: a package for downloading and preprocessing microdata from Brazilian Health Informatics Department (DATASUS)

Microdatasus: paquete para descarga y preprocesamiento de microdatos del Departamento de Informática del SUS (DATASUS)

Raphael de Freitas Saldanha [1]
Ronaldo Rocha Bastos [2]
Christovam Barcellos [1]

doi: 10.1590/0102-311X00032419

Resumo

O objetivo do estudo foi desenvolver um algoritmo capaz de realizar o download *e o pré-processamento de microdados fornecidos pelo Departamento de Informática do SUS (DATASUS) para diversos sistemas de informações em saúde para a linguagem de programação estatística R. O pacote desenvolvido permite o* download *e o pré-processamento de dados de diversos sistemas de informação em saúde, com a inclusão da rotulagem dos campos categóricos nos arquivos. A função de* download *foi capaz de acessar diretamente e reduzir o volume de trabalho para a seleção de arquivos e variáveis de microdados junto ao DATASUS. Já a função de pré-processamento foi capaz de efetuar a codificação automática de diversos campos categóricos. Dessa forma, a utilização desse pacote possibilita um fluxo de trabalho contínuo no mesmo programa, no qual esse algoritmo permite o* download *e o pré-processamento, e outros pacotes do R permitem a análise de dados dos sistemas de informação em saúde do Sistema Único de Saúde (SUS).*

Software; Processamento Eletrônico de Dados; Sistemas de Informação em Saúde

Correspondência
R. F. Saldanha
Instituto de Comunicação e Informação Científica e Tecnológica em Saúde, Fundação Oswaldo Cruz.
Av. Brasil 4365, Pavilhão Haity Moussatché, Rio de Janeiro, RJ 21040-900, Brasil.
raphael.saldanha@icict.fiocruz.br

[1] Instituto de Comunicação e Informação Científica e Tecnológica em Saúde, Fundação Oswaldo Cruz, Rio de Janeiro, Brasil.
[2] Departamento de Estatística, Universidade Federal de Juiz de Fora, Juiz de Fora, Brasil.

Introdução

A melhoria dos sistemas de saúde e de seus processos de decisão depende fortemente da produção de dados sobre o seu funcionamento [1]. No Brasil, após a *Constituição Federal* de 1988 ter estabelecido o Sistema Único de Saúde (SUS), foi criado o Departamento de Informática do SUS (DATASUS) em 1991 visando à coleta e organização de dados referentes ao SUS [2].

Os sistemas de informação em saúde mantidos pelo DATASUS, ou em colaboração com ele, cobrem diversos aspectos da saúde populacional. Alguns são de natureza epidemiológica, como o Sistema de Informações sobre Mortalidade (SIM) e o Sistema de Informações sobre Nascidos Vivos (SINASC), que utilizam dados dos cartórios. Já outros sistemas têm objetivos administrativos, como o Sistema de Internações Hospitalares (SIH) e o Sistema de Informações Ambulatoriais (SIA), usando dados provenientes diretamente da assistência à saúde. Mesmo os sistemas de informação de origem administrativa e financeira contêm dados relevantes acerca da situação de saúde brasileira [3].

A disseminação desses dados é realizada pelo DATASUS por meio de duas interfaces oferecidas aos usuários: TabNet e TabWin [4]. O TabNet é uma interface de produção de tabelas de dados agregados por meio do acesso a microdados contidos nos seus servidores de dados. Ele também possibilita a consulta de dados e indicadores de diferentes sistemas de informação em saúde agregados em unidades de tempo ou unidades geográficas.

O TabWin é uma interface de acesso do tipo "cliente", que permite a leitura dos arquivos de microdados do DATASUS disponibilizados no formato DBC. A utilização dos microdados anonimizados disponibilizados pelo DATASUS permite a realização de pesquisas com maior flexibilidade e detalhamento por parte do usuário por não estarem agregados em unidades preestabelecidas de tempo ou região.

Contudo, o TabWin apresenta algumas limitações [4,5]. Dentre elas, pode-se mencionar que ele pode ser executado apenas em um sistema operacional (Microsoft Windows) e não oferece a opção de *download* direto dos microdados, que devem ser baixados e organizados previamente pelo usuário. Além disso, a análise de dados no TabWin é limitada e precisa ser realizada em pacotes estatísticos dedicados.

Como alternativa à interface ao TabWin e superando algumas de suas limitações, este trabalho desenvolveu e disponibilizou um pacote para o programa estatístico R (http://www.r-project.org), com funções de *download* e pré-processamento de microdados do DATASUS. Ainda que o R apresente uma curva de aprendizado árdua, considera-se que esta dificuldade inicial é recompensada pela gama de possibilidades de manipulação e análise de dados que o programa permite.

Métodos

O programa estatístico R é uma linguagem de programação versátil, que permite desde a manipulação de dados à sua análise por meio de métodos estatísticos.

A quantidade e finalidade dos comandos ou funções do R podem ser livremente ampliadas pela criação de novos pacotes de funções. O algoritmo desenvolvido foi disponibilizado num pacote denominado *microdatasus*, seguindo as recomendações de desenvolvimento de pacotes [6]. Ele oferece funções para o download e o pré-processamento de microdados disponibilizados pelo DATASUS.

Instalação do pacote

O pacote pode ser instalado por meio de seu repositório no *website* GitHub (https://github.com/). Com o pacote *devtools* previamente instalado no R, o pacote *microdatasus* pode ser instalado da seguinte forma:

```
devtools::install_github("rfsaldanha/microdatasus")
```

Função *fetch_datasus*

Para o acesso, *download* e leitura dos arquivos de microdados no formato DBC foi desenvolvida a função *fetch_datasus*. Após especificar à função qual sistema de informação em saúde e qual a cobertura no tempo desejada, ela realiza o *download* dos respectivos microdados disponíveis no *website* do DATASUS (http://datasus.saude.gov.br/). A importação dos dados é realizada utilizando-se, interna mente, o pacote *read.dbc* (https://CRAN.R-project.org/package=read.dbc).

A função apresenta a seguinte estrutura básica:

```
fetch_datasus(year_start, month_start, year_end, month_end,
              uf = "all", information_system, vars = NULL)
```

Onde *year_start* é o ano do início da cobertura dos dados; *month_start* é o mês do início da cobertura dos dados; *year_end* é o ano do fim da cobertura dos dados; *month_end* é o mês do fim da cobertura dos dados; *uf* é a lista das Unidades Federativas que devem ser baixadas, informadas por intermédio de suas siglas oficiais [p.ex.: uf = c("RJ", "MG", "SP")]; *information_system* é a sigla do sistema de informação de saúde que deve ser acessado; e *vars* é a lista de variáveis que deve ser mantida após o *download*. Por padrão, a função mantém todas as variáveis encontradas nos microdados.

Em sua versão atual, a função *fetch_datasus* permite o *download* e a leitura dos arquivos de microdados dos sistemas SIM, SINASC, SIH, CNES (Cadastro Nacional dos Estabelecimentos de Saúde) e SIA, conforme o Quadro 1.

Cabe destacar que os argumentos relativos à época de cobertura dos dados (*year_start*, *month_start*, *year_end* e *month_end*) são referentes aos anos e meses de processamento dos casos pelo DATASUS. Por exemplo, óbitos ocorridos em dezembro de 2017 podem ser encontrados em arquivos de dezembro de 2017 e janeiro de 2018.

Os argumentos *month_start* e *month_end* são aplicados no *download* de dados dos sistemas de informação cujos arquivos de microdados são mensais, conforme Quadro 1.

A utilização do argumento *vars* pode ser de interesse quando pretende-se aplicar a função em um grande período temporal ou grande abrangência regional de cobertura, o que resulta em um grande número de registros. Para poupar recursos computacionais nesse caso, pode-se limitar o número de variáveis de interesse por meio do argumento *vars*, reduzindo o tamanho final do *data.frame*.

Cabe destacar que o *download* dos arquivos de microdados são realizados para uma pasta temporária do R no computador cliente e eliminados após a execução da função.

Funções de pré-processamento

Foram criadas funções de pré-processamento específicas para cada sistema de informação em saúde. Essas funções realizam o tratamento dos campos de acordo com o sistema de informação em saúde obtido, convertendo cada variável ao seu formato correto (texto, número inteiro, número decimal ou categórico) e imputando rótulos dos campos categóricos.

A versão atual do pacote permite o pré-processamento do SIM (todas as subdivisões), SINASC e SIH-RD, conforme o Quadro 2.

Os argumentos dessas funções são: *data* (objeto criado com a função *fetch_datasus*); *municipality_data* (enriquecimento dos dados referentes ao município de residência); e *information_system* (sistema de informações em saúde específico).

O argumento *municipality_data* acrescenta informações referentes ao município de residência do caso, como o nome do município, nome da Unidade Federativa, latitude, longitude da sede administrativa e área territorial.

O argumento *information_system* existe apenas na função dedicada a dados do SIH. Atualmente, essa função suporta apenas dados do SIH-RD, e toma esse argumento como padrão. Futuramente, a função será expandida para suportar arquivos de outros subsistemas do SIH.

A rotulagem dos campos categóricos é a tarefa principal dessa função. As informações referentes aos códigos dos campos categóricos foram obtidas baseando-se nos arquivos de definição (extensão

Quadro 1

Sistemas de informação, subdivisões, siglas e períodos de cobertura.

SISTEMA	SUBDIVISÃO	SIGLA	COBERTURA	MENSAL
SIH	AIH reduzida	SIH-RD	1992 – atual	Sim
	AIH rejeitada	SIH-RJ	1992 – atual	Sim
	AIH serviços profissionais	SIH-SP	1992 – atual	Sim
	AIH rejeitadas com código de erro	SIH-ER	1992 – atual	Sim
SIM	Declarações de óbitos	SIM-DO	1979 – atual	Não
	Declarações de óbitos fetais	SIM-DOFET	1979 – atual	Não
	Declarações de óbitos por causas externas	SIM-DOEXT	1979 – atual	Não
	Declarações de óbitos infantis	SIM-DOINF	1979 – atual	Não
	Declarações de óbitos maternos	SIM-DOMAT	1996 – atual	Não
SINASC	Declarações de nascidos vivos	SINASC	1994 – atual	Não
CNES	Leitos	CNES-LT	Outubro/2005 – atual	Sim
	Estabelecimentos	CNES-ST	Agosto/2005 – atual	Sim
	Dados complementares	CNES-EQ	Agosto/2005 – atual	Sim
	Equipamentos	CNES-EQ	Agosto/2005 – atual	Sim
	Serviço especializado	CNES-SR	Agosto/2005 – atual	Sim
	Habilitação	CNES-HB	Março/2007 – atual	Sim
	Profissional	CNES-PF	Agosto/2005 – atual	Sim
	Equipes	CNES-EP	Abril/2007 – atual	Sim
	Regra contratual	CNES-RC	Março/2007 – atual	Sim
	Incentivos	CNES-IN	Novembro/ 2007 – atual	Sim
	Estabelecimentos de ensino	CNES-EE	Março/2007 – atual	Sim
	Estabelecimento filantrópico	CNES-EF	Março/2007 – atual	Sim
	Gestão e metas	CNES-GM	Junho/2007 – atual	Sim
SIA	APAC de acompanhamento à cirurgia bariátrica	SIA-AB	Janeiro/2008 – Março/2013	Sim
	APAC de acompanhamento pós-cirurgia bariátrica	SIA-ABO	Abril/2013 – atual	Sim
	APAC de confecção de fístula arteriovenosa	SIA-ACF	Junho/2014 – atual	Sim
	APAC de laudos diversos	SIA-AD	Janeiro/2008 – atual	Sim
	APAC de medicamentos	SIA-AM	Janeiro/2008 – atual	Sim
	APAC de nefrologia	SIA-AN	Janeiro/2008 – Outubro/2014	Sim
	APAC de quimioterapia	SIA-AQ	Janeiro/2008 – atual	Sim
	APAC de radioterapia	SIA-AR	Janeiro/2008 – atual	Sim
	APAC de tratamento dialítico	SIA-ATD	Junho/2014 – atual	Sim
	Produção ambulatorial	SIA-PA	Junho/1994 – atual	Sim
	Psicossocial	SIA-PS	Janeiro/2013 – atual	Sim
	Atenção domiciliar	SIA-SAD	Novembro/2012 – atual	Sim

AIH: Autorização de Internação Hospitalar; APAC: Autorização de Procedimento de Alta Complexidade; CNES: Cadastro Nacional dos Estabelecimentos de Saúde; SIA: Sistema de Informações Ambulatoriais; SIH: Sistema de Internações Hospitalares; SIM: Sistema de Informações sobre Mortalidade; SINASC: Sistema de Informações sobre Nascidos Vivos.

Fonte: elaborado pelos autores com base em informações constantes no website do DATASUS (http://datasus.saude.gov.br/).

Quadro 2

Sistemas de informação e funções de pré-processamento.

SISTEMA	FUNÇÃO DE PRÉ-PROCESSAMENTO
SIM	`process_sim(data, municipality_data = TRUE)`
SINASC	`process_sinasc(data, municipality_data = TRUE)`
SIH	`process_sih(data, information_system = "SIH-RD", municipality_data = TRUE)`

SIH: Sistema de Internações Hospitalares; SIM: Sistema de Informações sobre Mortalidade; SINASC: Sistema de Informações sobre Nascidos Vivos.
Fonte: elaboração própria.

DEF) criados para o aplicativo TabWin. Destaca-se que algumas variáveis de alguns sistemas não são contempladas por esses arquivos de definição do TabWin, não permitindo assim a rotulagem dessas variáveis.

Os esquemas de rotulagem completos estão documentados detalhadamente na página do projeto, acessível em https://github.com/rfsaldanha/microdatasus.

Resultados

A utilização das duas funções contempladas no pacote *microdatasus* permitiu realizar o *download* e o pré-processamento conforme observa-se:

Consulta ao SIM-DO, ano de 2014 para todas as Unidades Federativas

Download dos dados:

```
> dados_brutos <- fetch_datasus(year_start = 2014, year_end = 2014,
                                information_system = "SIM-DO")
```

Pré-processamento:

```
> dados <- process_sim(data = dados_brutos)
```

Resultados:

```
> dim(dados)
[1] 1227039    103

> table(dados$SEXO)

 Feminino Masculino
  532.362   693.922

> prop.table[table(dados$RACACOR, useNA = "ifany")]*100

   Amarela     Branca  Indígena      Parda     Preta      <NA>
 0.5507567 51.2397731 0.2963231 35.4251984 7.5520012 4.9359474
```

Foram encontradas 1.227.039 declarações de óbitos no Brasil para o período, compondo um banco de dados com 103 variáveis. Dessas declarações, 532.362 óbitos foram do sexo feminino e 693.922 do sexo masculino. Sobre raça/cor, 51,24% dos óbitos foram identificados como de pessoas da cor branca e 4,93% não continham informações.

SIM-DO, 2013 a 2014, campos específicos, Estado do Rio de Janeiro

Aquisição dos dados:

```
dados_brutos <- fetch_datasus(year_start = 2013, year_end = 2014,
                    information_system = "SIM-DO",
                    vars = c("CODMUNRES", "DTOBITO", "CAUSABAS"),
                    uf = c("RJ"))
```

Pré-processamento:

```
dados <- process_sim(data = dados_brutos)
```

Resultados:

```
> dim(dados)
[1] 261076     11

> table(dados$SEXO)

 Feminino  Masculino
  559.956    587.510
```

Foram encontradas 261.076 declarações de óbitos para o Estado do Rio de Janeiro no período especificado. Além das três variáveis selecionadas na consulta, outras oito variáveis foram adicionadas sobre o município de residência com a função *process_sim*, em que o argumento *municipality_data* é verdadeiro por padrão.

SINASC, 2013, UFs específicas

Aquisição dos dados:

```
> dados_brutos <- fetch_datasus(year_start = 2013, year_end = 2013,
                              information_system = "SINASC",
                              uf = c("RJ", "SP", "MG", "ES")
```

Pré-processamento:

```
> dados <- process_sinasc(data = dados_brutos)
```

Resultados:

```
> dim(dados)
[1] 1147627      70
> prop.table[table(dados$LOCNASC, useNA = "ifany")]*100

                     Domicílio                       Hospital
                   0.227774355                   99.431696884
Outro estabelecimento de saúde                        Outros
                   0.248774210                    0.088094825
                          <NA>
                   0.003659726
```

Foram encontrados 1.147.627 declarações de nascidos vivos na Região Sudeste para os arquivos processados no ano de 2013. Nota-se que 99,43% desses nascimentos ocorreram em hospitais.

SIH-RD, 1º semestre de 2014, sem campos adicionais

Aquisição dos dados:

```
> dados_brutos <- fetch_datasus(year_start = 2014, month_start = 1,
                              year_end = 2014, month_end = 6,
                              information_system = "SIH-RD")
```

Pré-processamento:

```
> dados <- process_sih(data = dados_brutos,
                      information_system = "SIH-RD",
                      municipality_data = FALSE)
```

Resultados:

```
> dim(dados)
[1] 5722339     113
```

```
> prop.table[table(dados$COMPLEX, useNA = "ifany")]*100
```

```
 Alta complexidade Média complexidade
          6.373635          93.626365
```

Foram encontradas 5.722.339 Autorizações de Internação Hospitalar (AIH) no Brasil para o primeiro semestre de 2014. Dessas, 6,37% foram classificadas como de alta complexidade.

Discussão

A criação deste pacote tornou possível o acesso aos dados dos sistemas de informação em saúde mantidos pelo DATASUS diretamente por meio do programa R. Isso torna possível a adoção de um fluxo de trabalho linear, sem a necessidade de utilização de diferentes programas para aquisição, pré-processamento e análise de dados, dando celeridade e otimizando o processo de trabalho e a organização dos dados pelo usuário.

Políticas sobre dados abertos em entidades governamentais estão sendo amplamente discutidas atualmente e pode-se afirmar que o Ministério da Saúde, por intermédio do DATASUS, detém ampla experiência neste quesito, demonstrando um inegável pioneirismo na disseminação de dados em um sistema de cobertura universal.

Melhorias e modernizações são possíveis nos mecanismos de disseminação desses dados [5], como na documentação dos arquivos, metadados, estratégias de versionamento e melhor distinção entre data de processamento e real data do evento. Isso vem reforçar a necessidade de manutenção e investimento nos sistemas de informação em saúde para o conhecimento amplo e incondicional das condições de saúde da população brasileira.

Colaboradores

R. F. Saldanha contribuiu com a concepção e delineamento do artigo, aplicação do método e redação do manuscrito. R. R. Bastos e C. Barcellos contribuíram com o delineamento do artigo e sua revisão crítica.

Informações adicionais

ORCID: Raphael de Freitas Saldanha (0000-0003-0652-8466); Ronaldo Rocha Bastos (0000-0001-9597-5967); Christovam Barcellos (0000-0002-1161-2753).

Agradecimentos

Este artigo é parte integrante da tese de doutorado vinculada ao Programa de Pós-graduação Stricto Sensu em Informação e Comunicação em Saúde (PPGICS), do Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (Icict), da Fundação Oswaldo Cruz (Fiocruz), intitulada *Da Aquisição a Visualização de Dados: Aplicações do Processo KDD em Saúde*. O presente trabalho foi em parte financiado pela Fiocruz.

Referências

1. Handley K, Boerma T, Victora C, Evans TG. An inflection point for country health data. Lancet Glob Health 2015; 3:e437-8.
2. Ministério da Saúde. DATASUS: trajetória 1991-2002. Brasília: Ministério da Saúde; 2002.
3. Bittencourt SA, Camacho LAB, Leal MC. O Sistema de Informação Hospitalar e sua aplicação na saúde coletiva. Cad Saúde Pública 2006; 22:19-30.
4. Silva NP. A utilização dos programas TABWIN e TABNET como ferramentas de apoio à disseminação das informações em saúde [Dissertação de Mestrado]. Rio de Janeiro: Escola Nacional de Saúde Pública Sergio Arouca, Fundação Oswaldo Cruz; 2009.
5. Jorge MHPM, Laurenti R, Gotlieb SLD. Avaliação dos sistemas de informação em saúde no Brasil. Cad Saúde Colet (Rio J.) 2010; 18:7-18.
6. Wickham H. R packages: organize, test, document, and share your code. Sebastopol: O'Reilly; 2015.

Abstract

This study aimed to develop an algorithm for downloading and preprocessing microdata furnished by the Brazilian Health Informatics Department (DATASUS) for various health information systems, using the R statistical programming language. The package allows downloading and preprocessing data from various health information systems, with the inclusion of labeling categorical fields in the files. The download function was capable of directly accessing and reducing the workload for the selection of microdata files and variables in DATASUS, while the preprocessing function enabled automatic coding of various categorical fields. The package thus enables a continuous workflow in the same program, in which the algorithm allows downloading and preprocessing and other packages in R allow analyzing data from the health information systems in the Brazilian Unified National Health System (SUS).

Software; Electronic Data Processing; Health Information Systems

Resumen

El objetivo del estudio fue desarrollar un algoritmo capaz de realizar la descarga y pre-procesamiento de microdatos, proporcionados por el Departamento de Informática del SUS (DATASUS), para diversos sistemas de información en salud, así como para el lenguaje de programación estadístico R. El paquete desarrollado permite la descarga y preprocesamiento de datos de diversos sistemas de información en salud, con la inclusión del rótulo de los campos categóricos en los archivos. La función de descarga se mostró capaz de acceder directamente y reducir el volumen de trabajo para la selección de archivos y variables de microdatos a través del DATASUS, mientras que la función de pre-procesamiento fue capaz de efectuar la codificación automática de diversos campos categóricos. De esta forma, la utilización de este paquete posibilita un flujo de trabajo continuo en el mismo programa, donde este algoritmo permite la descarga y preprocesamiento y otros paquetes del R permiten el análisis de datos de los sistemas de información en salud del Sistema Único de Salud (SUS).

Programas Informáticos; Procesamiento Automatizado de Datos; Sistemas de Información en Salud

Recebido em 20/Fev/2019
Versão final reapresentada em 01/Mai/2019
Aprovado em 17/Jun/2019

RESEARCH

Dataset of hospital admissions authorizations in Brazil

Raphael F Saldanha*, Marcel M Pedroso, Rebecca Pontes Salles, Igor S Morais, Balthazar Paixão, Lucas Z Carraro, Carlos A M de Sousa, Pedro Teixeira and Jefferson C Lima

*Correspondence:
raphael.saldanha@icict.fiocruz.br
Fundação Oswaldo Cruz, Instituto de Comunicação e Informação Científica e Tecnológica em Saúde, Laboratório de Informação em Saúde, Plataforma de Ciência de Dados aplicada à Saúde, Av. Brasil, 4036 - Pavilhão da Expansão da Fiocruz, sala 713 - Manguinhos, 21040-361 Rio de Janeiro, RJ, Brazil
Full list of author information is available at the end of the article

Abstract

Objectives: Hospital admissions data is an important source of information for health surveillance and health policy makers. The Brazilian Health Ministry offers access to this data in scattered files using a legacy format, justifying the effort of this work to present this data in a unique, coherent and comprehensible format and infrastructure, suitable to its big data dimension and importance.

Data description: The produced dataset covers hospital admissions in the Brazilian Universal Health System as a whole, keeping the characteristics of the original files and with new variables added with correction results and enrichments.

Keywords: Health informatics; Hospitalization/statistics & numerical data; author; Data Science

Objective

Hospital admissions data is an important source of information for health surveillance and health policy makers [1, 2]. It enables many applications including the use of machine learning algorithms to predict several outcomes [3]. The Brazilian Health Ministry, through its Department of Health Informatics ("DataSUS"), publish monthly anonymized data files about hospital admissions in the Brazilian Universal Health System ("SUS") covering the Hospital Information System, called "Sistema de Informações Hospitalares" [4].

The lines in the files represent hospital admission authorizations ("Autorização de Internação Hospitalar" - AIH), where the circumstance of the admission is detailed. Main disease cause and procedure of treatment are stated, alongside with several other details. This data is made publicly available using a legacy file format called DBC, a compressed format derived from the DBF format, being published with a 2 or 3 months delay. The files are divided by year, month and state where the hospital admission was processed. Considering that Brazil have 27 states ("Unidades Federativas"), there are 324 files for each year. The public DataSUS File Transfer Protocol (FTP) has data files covering hospital admissions since 1992. Through the years, there has been several modifications in the number and structure of the collected variables, being more stable and compatible after 2008 [5].

Having the data scattered in many files using a legacy format imposes several difficulties for research, including challenges for file importation and combining procedures. Moreover, the categorical variables are coded and the data dictionary is inconclusive for some cases, making necessary a specialized knowledge of the data.

These are some of the challenges that justify this effort to provide the same data in a unique, coherent and comprehensible format and infrastructure that is suitable to its *big data* dimension and importance.

Data description

The database is created and maintained by the team of the Platform of Data Science Applied to Health ("Plataforma de Ciência de Dados Aplicada à Saúde" - PCDaS). The PCDaS is a research group from the Health Information Laboratory ("Laboratório de Informação em Saúde" - LIS), in the Institute of Communication and Scientific and Technological Information in Health ("Instituto de Comunicação e Informação Científica e Tecnológica em Saúde" - ICICT) of the Oswaldo Cruz Foundation ("Fundação Oswaldo Cruz" - Fiocruz) in partnership with the National Laboratory for Scientific Computing ("Laboratório Nacional de Computação Científica" - LNCC). Its main objective is to offer technological services and scientific computation to storage, management and analysis of large amounts of data for researchers, teachers and students from educational and research institutions.

The PCDaS offers a computational structure hosted in the LNCC facilities that was used for the Extraction, Transform and Load process (ETL) that produced this dataset. The ETL process was conducted using the Dataiku Data Science Studio (DSS) software (free edition) [6] as an ETL platform, alongside with R and Python scripts, as detailed bellow.

Extration

The original DBC data files are provided by DataSUS in a public FTP address with the addition of a data dictionary. They are downloaded by an R script in the beginning of the ETL process.

Transform

After downloading the data of all states for one year and month, the DBC files are converted to a DBF format and loaded using an *R* script. They are combined into one data frame covering all AIHs for one month.

This data frame is pre-processed and enriched using both DSS pre-implemented processors and other customized functions implemented in Python. The categorical variables are labelled, date variables are formatted, additional information about the municipalities of residence of the patient and treatment are included and additional information about the hospitalization process, diagnosis and clinical procedures are created. This involved an extensive research about the variables, the data dictionaries available and other documents created by municipal and state level health secretaries.

Load

The processed and enriched data are loaded into an ElasticSearch [7] index using a Python script and also exported to a CSV file. ElasticSearch was choose by its capabilities to deal with large amounts of data and a its stack of other tools, like Kibana for visualization.

The ETL process described is repeated monthly, beginning with the 2008 files. Moreover, the same process is re-executed as new data is made available in the DataSUS public FTP address.

Encompassing from 2008 to 2018, the dataset contains 127.402.174 records. The original DBC files contained 113 variables, all named in capital letters. The enrichment process adds 127 new variables, named in lowercase, resulting in 240 variables in total.

Checking

After the loading process, the integrity of the data was tested by comparing the counts of records by state and year with the official numbers of hospital admissions accessible at the TabNET website.

Dissemination

The resulting ElasticSearch index and the CSV files are made publicly available and accessible in a CKAN repository. Moreover, this repository provides access to several useful resources produced by the team of PCDaS including a detailed data dictionary, the fully documented ETL process, an interactive visualization dashboard powered by Kibana [7] and a tutorial for using the ElasticSearch index for data mining. The repository can be accessed with this link: `https://bigdata-metadados.icict.fiocruz.br/dataset`

Table 1 Overview of data files

Label	Name of data file/data set	File types	DOI
Dataset 1	DataSUS SIH	ElasticSearch index	000
Dataset 2	DataSUS SIH	CSV extension	000

Limitations

Since missing information were not treated, some known limitations about the quality of the original data [cite] are persistent on the published pre-processed and enriched data. However, this limitation was expected, and it is in conformity with the objective of keeping the data comparable with its source.

Nonetheless, possibly erroneous information were treated when possible, resulting in the creation of new variables, therefore enabling the comparison between the original data field and the pre-processed one.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

RFS was responsible for the ETL process, data maintenance and wrote the manuscript with the support of all authors. ISM and LZC were responsible for the computational infrastructure. RPS, BSCP, CAMS and PT contributed to the design and implementation of the research. MMP and JCL were in charge of overall direction and planning. All authors provided critical feedback and helped shape the research, dataset production and manuscript.

Acknowledgements

We thank the Fundação Oswaldo Cruz (Fiocruz), Laboratório Nacional de Computação Científica (LNCC), Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) and Sistema Único de Saúde (SUS) for supporting this research.

References

1. Lessa, F.J.D., Mendes, A.d.C.G., Farias, S.F., de Sá, D.A., Duarte, P.O., de Melo Filho, D.A.: Novas metodologias para vigilância epidemiológica: uso do Sistema de Informações Hospitalares - SIH/SUS. Informe Epidemiológico do Sus **9**, 3–27 (2000). doi:10.5123/S0104-16732000000500001
2. Bittencourt, S.A., Camacho, L.A.B., Leal, M.d.C.: O Sistema de Informação Hospitalar e sua aplicação na saúde coletiva. Cadernos de Saúde Pública **22**(1), 19–30 (2006). doi:10.1590/S0102-311X2006000100003
3. Chiavegatto-Filho, A.D.P.: Uso de big data em saúde no Brasil: perspectivas para um futuro próximo. Epidemiologia e Serviços de Saúde **24**(2), 325–332 (2015). doi:10.5123/S1679-49742015000200015
4. MS: DataSUS (2016). `http://www2.datasus.gov.br` Accessed 2015-04-20
5. Brasil: DataSUS - Trajetória 1991-2002. Technical report, Ministério da Saúde, Brasília (2002)
6. Dataiku: Dataiku tool (2020). `https://www.dataiku.com/` Accessed 2020-12-10
7. Elastic: Elastic Search (2020). `https://www.elastic.co/` Accessed 2020-12-10

4.3 Do constructo teórico à mineração

Nesta seção, é apresentado um artigo denominado "Estudo de análise de rede do fluxo de pacientes de câncer de mama no Brasil entre 2014 e 2016", tendo sua aprovação final em 21 de fevereiro de 2019 e publicado na Cadernos de Saúde Coletiva.

O artigo aborda etapas de leitura e transformação de dados relacionados ao fluxo de mama no Brasil, seguido de uma análise conduzida através de métodos de Análise de Redes sobre o fluxo de pacientes oncológicos no Brasil, com ênfase nas etapas de pré-processamento e transformação, levando a uma etapa de mineração de dados utilizando métodos de Análise de Redes.

CSP CADERNOS DE SAÚDE PÚBLICA
REPORTS IN PUBLIC HEALTH

ARTIGO
ARTICLE

Estudo de análise de rede do fluxo de pacientes de câncer de mama no Brasil entre 2014 e 2016

Analytical study of the breast cancer patient flow network in Brazil from 2014 to 2016

Estudio de análisis de red del flujo de pacientes con cáncer de mama en Brasil entre 2014 y 2016

Raphael de Freitas Saldanha [1]
Diego Ricardo Xavier [1]
Keila de Morais Carnavalli [1]
Kátia Lerner [1]
Christovam Barcellos [1]

doi: 10.1590/0102-311X00090918

Resumo

Este estudo busca analisar o fluxo de pacientes oncológicos de mama que são atendidos fora de seu domicílio de residência. Foram considerados as internações hospitalares e os tratamentos por quimioterapia e radioterapia para neoplasias malignas na mama, no âmbito do Sistema Único de Saúde, entre os anos de 2014 e 2016. Foi empregado o método de análise de redes, considerando o município de residência e de tratamento como nós de um grafo, que consiste em um "estudo de redes organizacionais de sistemas de saúde". Além disso, distância e tempo de deslocamento foram estimados por meio da melhor rota viável, segundo a malha rodoviária do projeto Open Street Maps*. Os resultados apontam que 51,34% dos pacientes de câncer de mama no Brasil foram atendidos fora de seu município de residência, seguindo fluxos que são regionalizados e que preservam fronteiras estaduais, em geral, em direção a capitais ou a cidades de grande porte. Por outro lado, os resultados também apontam exceções específicas, visto que alguns municípios detêm um grau de proeminência que supera os limites estaduais. O tempo de deslocamento entre município de residência e município de atendimento apresentou medianas próximas a três horas, e 75% dos deslocamentos se dão em até 324km para tratamento por quimioterapia, 287km para tratamento por radioterapia e 282km para internações. Esses resultados são indicativos das dificuldades de acesso aos serviços de oncologia, o que potencialmente agrava a experiência do adoecimento oncológico em termos de impacto no indivíduo e em sua família.*

Neoplasias de Mama; Acesso aos Serviços de Saúde; Sistemas de Informação

Correspondência
R. F. Saldanha
Laboratório de Informação e Saúde, Instituto de Comunicação e Informação Científica e Tecnológica em Saúde, Fundação Oswaldo Cruz.
Av. Brasil 4036, sala 210, Rio de Janeiro, RJ 21040-360, Brasil.
raphael.saldanha@icict.fiocruz.br

[1] Instituto de Comunicação e Informação Científica e Tecnológica em Saúde, Fundação Oswaldo Cruz, Rio de Janeiro, Brasil.

Introdução

Estima-se que o câncer seja responsável por 13% dos óbitos no mundo em 2008 [1], levando 7,6 milhões de pessoas [1] à morte anualmente, uma população de tamanho semelhante à da Cidade do México (México). Não obstante, estima-se que o número de casos de câncer aumente em 70% nas próximas duas décadas [1]. Os tipos de câncer mais incidentes no mundo são o de pulmão (1,8 milhão de casos), o de mama (1,7 milhão de casos), o de intestino (1,4 milhão de casos) e o de próstata (1,1 milhão de casos) [2].

O câncer configura-se no Brasil como um problema de saúde pública de dimensão nacional. A mortalidade por essa doença foi responsável por 16% dos óbitos – 3% a mais que a média mundial – segundo dados do Ministério da Saúde em 2014 (Departamento de Informática do SUS – DATASUS. http://www2.datasus.gov.br). De acordo com estimativas do Instituto Nacional de Câncer José Alencar Gomes da Silva (INCA), para o biênio 2018-2019, 600 mil novos casos surgirão a cada ano no Brasil [2].

Com mais de cem tipos existentes da doença, a redução da mortalidade no câncer pode ser efetivamente obtida por meio de diagnóstico precoce e rápido início do tratamento [3]. A Política Nacional de Prevenção e Controle de Câncer (*Portaria nº 874/2013* [4]) designa que o tratamento será realizado em estabelecimentos de saúde habilitados, como o Centro de Alta Complexidade em Oncologia (Cacon) ou Unidade de Alta Complexidade em Oncologia (Unacon). No Brasil, há 288 unidades e centros em oncologia cadastrados. A responsabilidade de organizar o atendimento e o encaminhamento do paciente portador de câncer fica a cargo das secretarias estaduais e municipais.

Um dos maiores problemas do atendimento aos pacientes tem sido a falta de uma rede consolidada de referência para o diagnóstico e o tratamento precoces de casos, o que pressupõe o acesso aos serviços em portas de entrada descentralizadas do sistema e o encaminhamento a unidades de tratamento próximas ao local de residência dos pacientes [5]. A distância e a presença de polos de atenção são, portanto, elementos-chave para o acesso aos serviços que, por sua vez, configuram regiões do ponto de vista funcional, bem como de seus fluxos internos [6].

A regionalização do sistema de saúde no país apresenta diversos fatores que tornam complexa a institucionalização de uma rede homogênea de serviços. Dentre esses, destacam-se a heterogeneidade territorial, a formalização de responsabilidades institucionais e a regulação centralizada com a manutenção da autonomia dos governos locais que, em última análise, é uma questão delicada em termos político-administrativos [7,8].

Do ponto de vista do acesso geográfico, alguns autores, por meio de técnicas distintas, realizaram trabalhos de mapeamento do fluxo de pacientes e identificação de redes de atendimento para tratamento de saúde [9,10,11,12]. Na prática, a distância dos deslocamentos populacionais em busca de atendimento aumenta à medida que aumenta a complexidade do serviço de saúde procurado. Alguns trabalhos apontam que essas redes de deslocamento de pacientes apresentam sobreposição em função da especialidade [10,11].

Considerando as necessidades apontadas de melhorias no ordenamento do território para estruturação do atendimento ao fluxo de pacientes oncológicos, o objetivo deste estudo é analisar o fluxo de pacientes para internação e tratamento de neoplasias malignas na mama, no âmbito do Sistema Único de Saúde (SUS), entre os anos de 2014 e 2016. Para isso, serão identificados os municípios de origem e de destino de pacientes para internações hospitalares, bem como para tratamento por quimioterapia e por radioterapia, conforme a metodologia capaz de descrever as principais estruturas da rede de atendimento desse tipo de câncer no país.

Métodos

Nas Ciências da Saúde, a metodologia de análise de redes supre a necessidade de utilização de técnicas específicas que possibilitem descrever e estudar as características iminentemente relacionais dos sistemas de saúde, comumente ignoradas por métodos que pressupõem a independência das unidades observadas.

Considerando a tipologia de estudos de redes descrita na obra de revisão de Luke & Harris [13], o presente trabalho se enquadra como um estudo da "estrutura interorganizacional de sistemas de

saúde", com o diferencial da aplicação de diversas métricas de modularidade e da formação de comunidades. Não são analisadas redes de interações sociais ou de transmissão de doenças, mas de entes organizacionais (neste caso, os municípios). Desse modo, é analisada uma rede formada por municípios que enviam e recebem pacientes. Os nós da rede são os municípios que enviam ou recebem pacientes, e as arestas entre os nós da rede são estabelecidas por meio desse fluxo de pacientes, ponderadas pela quantidade daqueles que foram enviados no período considerado.

A noção de redes e as metodologias específicas para sua análise têm atraído interesse considerável com a maior disponibilidade de dados relacionais advindos de redes sociais, sejam analógicas ou digitais [14,15]. De maneira geral, as metodologias voltadas ao estudo de redes buscam a detecção de padrões e regularidades entre os relacionamentos de unidades que interagem. Nessa perspectiva, a unidade de análise principal de redes está nas relações, e não somente nas unidades individuais e independentes, como nos métodos estatísticos e epidemiológicos clássicos [16,17].

Descrição dos dados

As bases de dados utilizadas para quantificação das internações e tratamento de pacientes de câncer no Brasil foram obtidas a partir do Sistema de Internações Hospitalares (SIH) e do Sistema de Informações Ambulatoriais (SIA) do DATASUS (http://www2.datasus.gov.br), para os anos de 2014 a 2016. Dessa forma, o escopo deste trabalho se limita aos pacientes tratados no SUS e em estabelecimentos de saúde conveniados.

As Autorizações de Internações Hospitalares (AIH) foram filtradas com base no diagnóstico principal de código "C50" da Classificação Internacional de Doenças – 10ª revisão (CID-10), referente à neoplasia maligna da mama. Os registros de AIH do tipo "5" (autorização de continuidade da internação) foram descartados.

Para os registros sobre procedimentos ambulatoriais de quimioterapia e radioterapia, foram utilizadas as bases de dados específicas das autorizações de procedimentos de alta complexidade (APAC) de quimioterapia e radioterapia, filtrando-se os registros com diagnóstico primário de código "C50".

Todos os dados foram obtidos junto ao DATASUS no formato DBC, convertidos para o formato DBF e tabulados no pacote estatístico R versão 3.4.1 (http://www.r-project.org). As tabelas de distribuição de frequências de município de origem e de destino criadas no R foram exportadas para o software Gephi versão 0.9.2 (https://gephi.org/users/download/), em que foram criados os sociogramas e foi realizado o cálculo das medidas descritivas das redes.

Com base na tabela de origem e destino dos pacientes de cada rede, foram obtidas estimativas das distâncias rodoviárias e do tempo de viagem entre os municípios, por meio do pacote OSRM versão 3.0.1 (https://github.com/Project-OSRM/node-osrm/issues/80). Essas medidas são baseadas na melhor rota possível entre as coordenadas da sede dos municípios de origem e destino, utilizando a malha rodoviária do projeto *Open Street Maps*.

O relacionamento entre os municípios é caracterizado como envio ou recebimento de pacientes para internação hospitalar e tratamento por quimioterapia ou radioterapia. Dessa maneira, os relacionamentos entre os municípios são direcionais e assimétricos (partindo do município de origem para o de destino), bem como ponderados pela quantidade de pacientes observados no período, em cada par de municípios. Cabe ressaltar que, além de enviar e receber pacientes, um mesmo município também recebe seus residentes em unidades de saúde (fluxo local). As medidas descritivas das redes foram calculadas apenas com base nos municípios que apresentam algum relacionamento, mesmo que seja com ele mesmo (fluxo local). Municípios que não enviaram e não receberam nenhum paciente no período foram descartados.

Foram arrolados como atributos básicos dos relacionamentos os seguintes: município de residência, município de tratamento, número de internações ou tratamentos no período, distância rodoviária entre os municípios e tempo estimado de viagem. Nos casos de coincidência entre município de origem e de destino, as estimativas de distância e tempo de viagem foram desconsideradas.

Análise da estrutura da rede

A metodologia proposta por Blanchet & James [18] para o estudo de redes de relacionamentos em sistemas de saúde de países em desenvolvimento foi adotada neste trabalho, segmentando a análise pela descrição dos dados, pela definição dos atores e por relacionamentos e, por fim, pela análise estrutural da rede.

Para caracterização descritiva da rede, algumas medidas globais (referentes à totalidade da rede) e individuais (para cada município) foram calculadas para as três redes. O Quadro 1 apresenta definições breves de algumas medidas aplicadas ao contexto desta pesquisa. Detalhes sobre o cálculo dessas medidas podem ser encontrados nas obras de Wasserman & Faust [16] e Valente [17].

Resultados

A Tabela 1 apresenta os resultados obtidos para as redes de internações, quimioterapia e radioterapia.

Conforme pode ser observado, entre 2014 e 2016, foram aprovadas 177.841 AIHs relacionadas a câncer de mama, 4.348.404 APACs de quimioterapia e 200.929 APACs de radioterapia. Desse total,

Quadro 1

Definições das medidas de análise de rede utilizadas.

Nível	Medida	Definição e interpretação
Global	Grau médio	O grau de um município corresponde à quantidade de municípios com o qual ele se relaciona, recebendo ou enviando pacientes. Esse grau pode ser ponderado pela quantidade de pacientes enviados ou recebidos durante o período. O grau médio representa a quantidade média de conexões que os municípios apresentam ou, quando ponderado, a quantidade média de internações ou procedimentos por município conectado.
	Diâmetro	A maior distância geodésica na rede, ou seja, a maior distância encontrada entre os pares de municípios. Diâmetros maiores podem significar uma rede menos interconectada.
	Modularidade	Mede a conexão de uma rede e a capacidade de se dividir em módulos ou comunidades. Quanto maior a modularidade, maior é a tendência da rede em apresentar grupos ou comunidades de municípios que detêm grande número de conexões entre si e pequeno número de conexões com outros grupos na rede.
Individual	Grau Grau de entrada Grau de saída	O grau individual de um município corresponde ao número de municípios com os quais ele se relaciona no envio ou recebimento de pacientes. O grau de entrada refere-se ao número de municípios dos quais ele recebe pacientes, ao passo que o grau de saída se refere ao número de municípios para os quais ele envia pacientes. Essas três medidas podem ser ponderadas pelo número de pacientes constantes nas relações.
	Autocentralidade	Mede o quanto um município tende a se relacionar com outros municípios com maiores graus de centralidade de proximidade. Uma maior medida de autocentralidade indica que um município tende a receber pacientes de municípios centrais da rede de atendimento.
	Comunidade	Considerando a modularidade da rede, são agrupados municípios que detêm intensa relação entre si em termos de conectividade e fraca relação com municípios de outros grupos. Nesse cálculo, foi adotado o método não supervisionado de Louvain (https://gephi.org/users/download/).

Tabela 1

Medidas e estatísticas das redes.

Medida/Rede	Internações	Quimioterapia	Radioterapia
Municípios conectados	5.098	4.729	4.729
Quantidade de internações ou procedimentos			
Total	177.841	4.348.404	200.929
Local	86.532	2.089.642	85.609
Não local	91.309	2.258.762	115.320
Grau médio	3,70 (15,48)	3,69 (18,25)	2,90 (14,21)
Grau médio de entrada	1,85 (15,24)	1,84 (17,95)	1,45 (14,11)
Grau médio de saída	1,85 (1,09)	1,84 (2,78)	1,45 (0,78)
Grau médio ponderado	69,03 (689,11)	1.612,86 (14.783,30)	84,65 (717,88)
Grau médio ponderado de entrada	34,52 (442,63)	806,43 (9.461,57)	42,33 (466,70)
Grau médio ponderado de saída	34,52 (256,39)	806,43 (5.528,87)	42,33 (265,28)
Diâmetro	13	12	8
Modularidade	0,88	0,906	0,923
Comunidades	40	37	39

Nota: entre parênteses, medida do desvio padrão.

48,66% das AIHs, 48,05% das APACs de quimioterapia e 42,61% das APACs de radioterapia foram atendidas no próprio município de residência. Assim, 52,15% dos pacientes de câncer de mama no Brasil foram atendidos fora de seu município de residência.

Em geral, os municípios tendem a se relacionar com três a quatro municípios para envio ou recebimento de pacientes; contudo, observa-se certa variabilidade nessa distribuição. As medidas de grau médio de saída e grau médio de entrada apresentaram concentração em valores inferiores, conforme indicado pelo coeficiente de assimetria positivo. Apenas a medida de grau médio de saída apresenta valores mais concentrados em torno da média. Esse comportamento sugere que boa parte dos municípios se relacionam com outros municípios para o envio de pacientes, e apenas alguns se relacionam para o recebimento destes. Essa tendência se assemelha ao se considerarem os graus ponderados. Na rede de AIHs, 75% dos municípios estabeleceram até 18 viagens por município conectado no período observado.

O diâmetro das redes variou entre 13 (AIH), 12 (APAC quimioterapia) e 8 (APAC radioterapia), indicando que a rede de radioterapia tende a ser mais coesa, ao passo que a rede de AIH tende a ser mais esparsa.

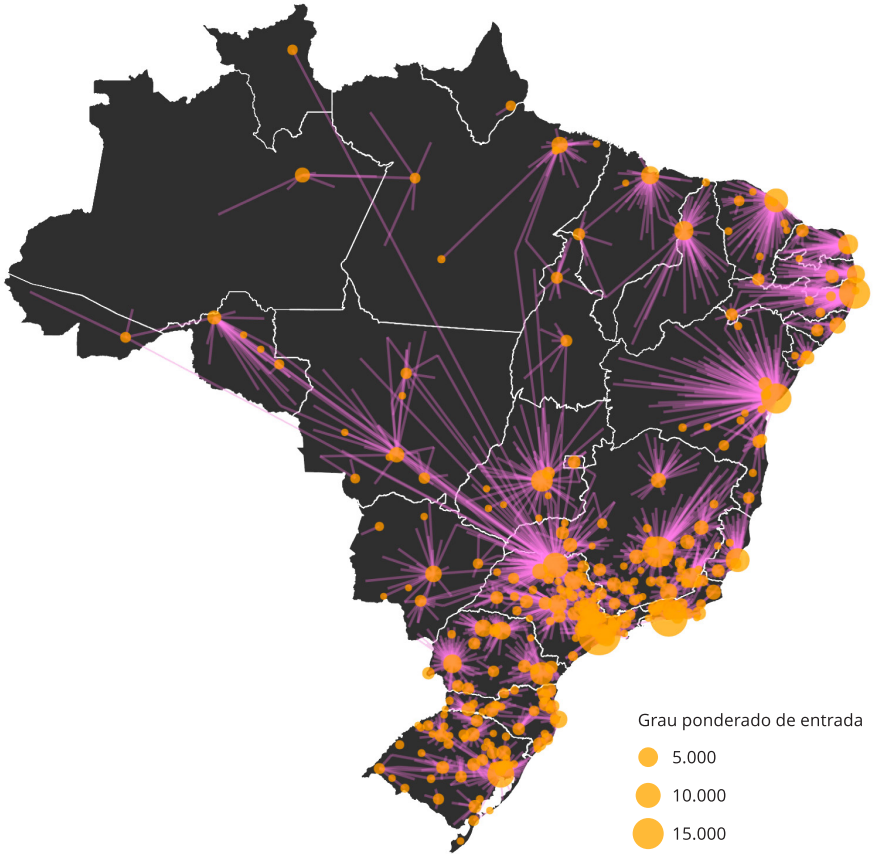
As três redes apresentaram altos valores de modularidade, o que indica uma tendência de formação de clusters de municípios para envio e recebimento de pacientes. O algoritmo identificou 40 comunidades na rede de AIHs, 37 comunidades na rede de APAC quimioterapia e 39 comunidades na rede de APAC radioterapia, o que pode ser considerado um número semelhante de comunidades, ainda que suas conformações possam ser diferentes.

As Figuras 1, 2 e 3 apresentam a distribuição espacial da rede de internações, de quimioterapia e de radioterapia, respectivamente. As linhas indicam a conexão entre os municípios, representados por círculos cujo diâmetro reflete proporcionalmente o grau ponderado de entrada.

Pode-se perceber que as conexões entre os municípios tendem a preservar as fronteiras estaduais, salvo algumas exceções, em geral em direção a capitais ou cidades de grande porte. Observa-se que a Região Nordeste tende a apresentar dinâmicas bem demarcadas pela divisão estadual, com um fluxo de cidades do interior para as capitais, com pouco ou nenhum fluxo entre as cidades do interior. Já nas outras regiões, como a Sudeste, o fluxo entre cidades do interior é presente. Na rede de internações para o Estado de Minas Gerais, por exemplo, um componente (um conjunto de nós integrados entre si, sem relacionamento com demais nós da rede) cujo município mais proeminente é Montes Claros não apresenta fluxo com a capital, Belo Horizonte. Percebe-se que a rede de quimioterapia é mais densa,

Figura 1

Mapa do fluxo de pacientes para internações. Brasil, 2014-2016.



apresentando mais fluxos interestaduais e interregionais que as demais redes, ao passo que as outras redes apresentam fluxos mais hierarquizados.

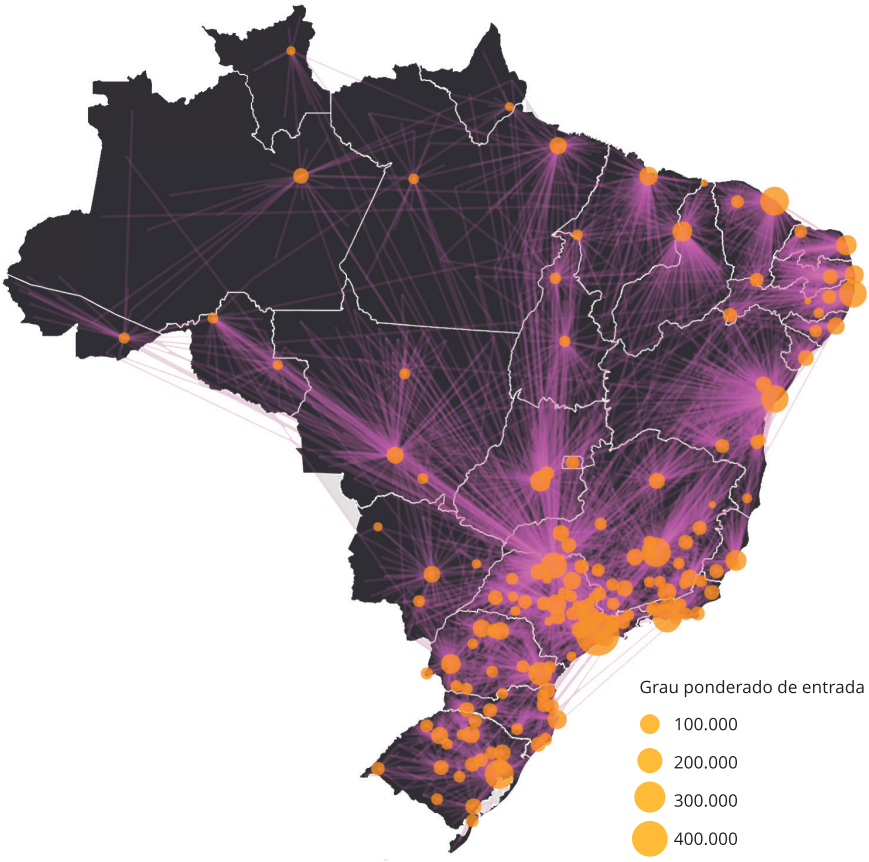
Ao se ordenarem os municípios pelo grau de entrada, Barretos (São Paulo) apresenta o maior grau de entrada nas três redes, recebendo pacientes de 617 municípios para internações, de 748 municípios para tratamentos por quimioterapia e de 540 municípios para tratamentos por radioterapia. Em seguida, nesse *ranking*, os dez primeiros resultados são ocupados por capitais estaduais. Ao se eliminarem as capitais, nota-se a proeminência dos municípios de Jaú (São Paulo), Cascavel (Paraná), Muriaé (Minas Gerais) e Campinas (São Paulo) por receberem pacientes de um grande número de municípios para internações hospitalares e tratamento.

Novamente, o Município de Barretos está em primeiro lugar no *ranking* das três redes, segundo a medida de autocentralidade. Ao se eliminarem as capitais desse ranqueamento, destaca-se também a proeminência dos municípios de Jaú, Campinas, São José do Rio Preto (São Paulo), Cascavel, Muriaé, Botucatu (São Paulo), Ribeirão Preto (São Paulo) e Passo Fundo (Rio Grande do Sul), dentre outros.

O tempo de deslocamento entre município de residência e município de atendimento apresentou medianas próximas a 3h nas três redes. Considerando o terceiro quartil desses tempos, a rede de

Figura 2

Mapa do fluxo de pacientes para quimioterapia. Brasil, 2014-2016.



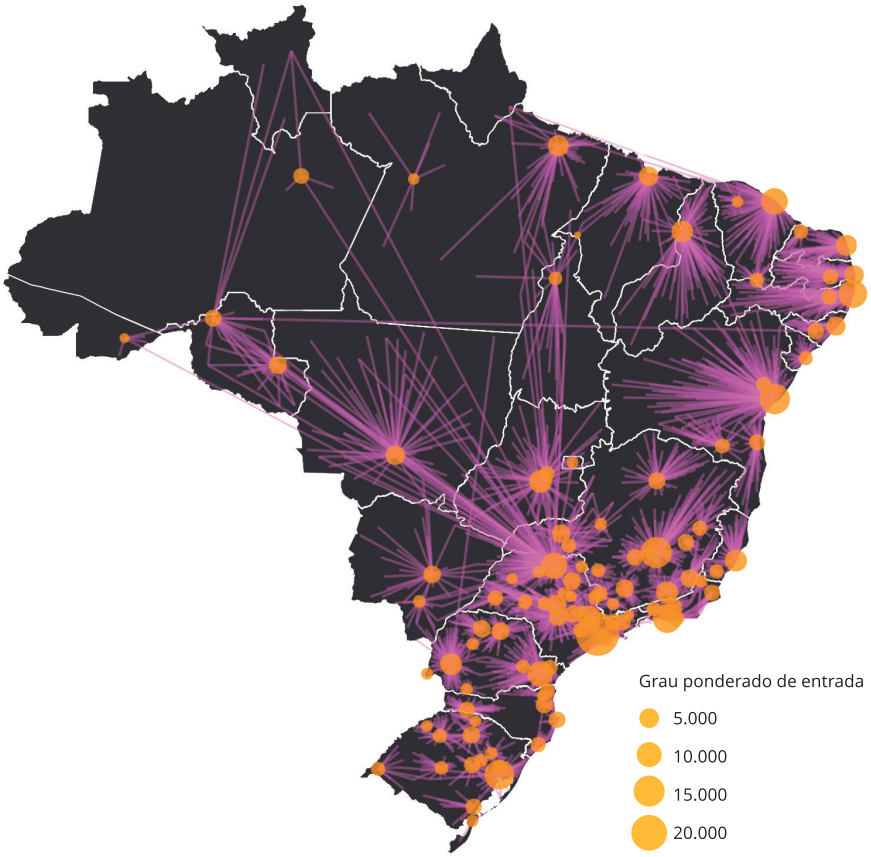
APAC quimioterapia apresentou o valor de 5h, ao passo que as outras redes apresentaram valores próximos a 4:30h.

Considerando as estimativas de distâncias percorridas entre os pares de município, pode-se notar que as três redes apresentam concentrações de valores inferiores. No entanto, ao se considerar o terceiro quartil das distribuições, revela-se que 75% dos pares percorrem até 324km para tratamento por quimioterapia, 287km para tratamento por radioterapia e 282km para internações.

A Figura 4 apresenta uma série de *boxplots* da estimativa da distância rodoviária percorrida, separada por unidade federativa de origem, para as internações hospitalares. Distâncias superiores a 3.000km foram desconsideradas. Observa-se que as maiores medianas de distância percorrida estão nos estados das regiões Norte e Nordeste. No Amapá, em Roraima e no Amazonas, mais de 50% das internações estão sujeitas a deslocamentos superiores a 500km. No Estado do Pará, 75% das internações hospitalares envolvem deslocamentos de até 1.000km. Já os estados das regiões Sudeste e Sul apresentam necessidades de deslocamento, em geral, inferiores a 250km.

Figura 3

Mapa do fluxo de pacientes para radioterapia. Brasil, 2014-2016.



Discussão

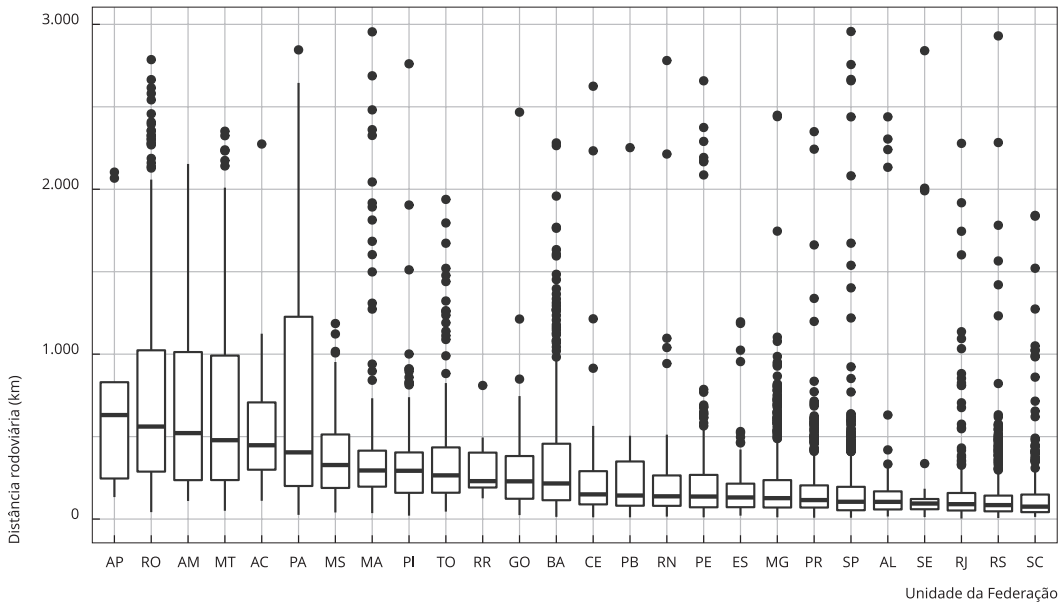
A experiência do adoecimento oncológico é agravada pelas dificuldades de acesso aos serviços de saúde, e se estende não só ao paciente, mas cria tensionamentos adicionais também aos familiares, aos amigos e aos profissionais de saúde [19]. Os resultados deste trabalho apontam que essa condição pode atingir 51,34% dos pacientes de câncer de mama no âmbito do SUS, impondo deslocamentos superiores a três horas de viagem na metade dos casos.

O impacto da doença sobre o indivíduo e a necessidade de manter vínculos com os serviços de saúde podem obrigar o paciente a se mudar, mesmo que temporariamente, para cidades com oferta e melhores condições de tratamento, provocando rupturas de redes sociais de apoio.

Este estudo utilizou uma metodologia de análise de redes que apresenta recursos para a caracterização e a predição de conjuntos de dados de diferentes naturezas. Pode-se afirmar que sua origem encontra-se na teoria matemática dos grafos. Contudo, se diferencia por sua natureza pragmática, já que envolve aspectos computacionais, gráficos e empíricos, além do universo teórico matemático [11,13]. É importante notar que trabalhos de outras áreas na saúde já têm adotado claramente a metodologia de análise de redes [20].

Figura 4

Boxplots da estimativa da distância rodoviária percorrida por Unidade da Federação de origem.



AC: Acre; AL: Alagoas; AM: Amazonas; AP: Amapá; BA: Bahia; CE: Ceara; ES: Espírito Santo; GO: Goiás; MA: Maranhão; MG: Minas Gerais; MS: Mato Grosso do Sul; MT: Mato Grosso; PA: Pará; PB: Paraíba; PE: Pernambuco; PI: Piauí; PR: Paraná; RJ: Rio de Janeiro; RN: Rio Grande do Norte; RO: Rondônia; RR: Roraima; RS: Rio Grande do Sul; SC: Santa Catarina; SE: Sergipe; SP: São Paulo; TO: Tocantins.

Constructos advindos da análise de redes têm sido aplicados em diferentes tipologias de populações para melhor compreender a disseminação de comportamentos, ideias ou mesmo doenças, além de estudar a efetividade de métodos de controle. De modo semelhante, esta metodologia pode ser aplicada efetivamente na análise da organização de sistemas de saúde [13,21].

O fluxo de pacientes de câncer no Brasil tem sido abordado por diferentes metodologias, como na análise de itinerários terapêuticos e na análise da oferta, do acesso e da demanda de tratamento [22,23]. Os trabalhos de Oliveira et al. [11] e Mancini [24] apresentam resultados específicos sobre o câncer de mama, ainda que se restrinjam a uma unidade federativa específica ou adotem metodologias diferenciadas. Esses e outros trabalhos adotam, em geral, distâncias euclidianas para medida de distância de deslocamento, ao passo que o presente estudo adota uma estimativa da distância rodoviária, buscando evidenciar de modo mais realístico as distâncias impostas aos pacientes que são atendidos fora de seu domicílio de residência.

O método de análise de redes aplicado no diagnóstico organizacional do fluxo de pacientes permitiu abordar novos aspectos sobre o atendimento de pacientes de câncer de mama no Brasil, medindo e comparando as redes de internações e tratamentos. A descrição dessas redes revela que elas assumem formatos diferenciados no território nacional, apresentando especificidades quanto a seu diâmetro e densidade.

Enquanto a medida de grau médio de entrada reflete a centralidade de um município de acordo com a quantidade de municípios dos quais ele recebe pacientes, a medida de autocentralidade destaca a centralidade de um município na rede, considerando o quanto ele se conecta com municípios

da rede que também apresentam uma relativa alta medida de centralidade [16]. Dessa maneira, esta última medida é capaz de revelar quais municípios apresentam um nível de especialização e referência na rede capaz de captar pacientes de outros grandes centros de internação e tratamento, como capitais estaduais.

As medidas de centralidade obtidas pelas capitais parecem demonstrar avanços em termos da regionalização e hierarquização dos serviços do SUS, pelo menos ao nível de Unidades da Federação (UF) [10]. Contudo, a medida específica de autocentralidade revela a proeminência de algumas cidades do interior dos estados, como Barretos), na captação de pacientes advindos de outros grandes centros de tratamento e internação, um comportamento que necessita ser avaliado à luz da regionalização preconizada pelos planos diretores estaduais de regionalização.

Especificamente, no caso de Barretos, a rede observada é composta de 540 municípios de diversos estados, o que sugere que a excelência de tratamento de câncer torna o município um centro de referência nacional. Possivelmente, a regionalização com implantação de centros de tratamento similares poderia tornar o acesso geográfico menos dispendioso aos usuários.

No Brasil, existem 438 regiões de saúde. Em alguns estados, foram conformadas mesorregiões de saúde, que se propõem a agrupar conjunto de regiões em busca de oferta de serviços com maior grau de complexidade. O presente estudo apontou 40 comunidades para internações, 37 para quimioterapia e 39 para radioterapia. Com base nos resultados da Figura 1, observa-se a conformação de, pelo menos, uma rede em cada UF. Contudo, no Sul e Sudeste do país, se concentra o maior volume de comunidades. Por um lado, esses resultados evidenciam a importância dos centros implantados nas capitais e a influência dos governos estaduais na oferta dos serviços. Por outro lado, impõem maior investimento no processo de descentralização de municípios com capacidade gerencial para implantação de centros de tratamento e possibilidade de acesso geográfico otimizado à população coberta.

A quantidade proporcional de pacientes que precisam se deslocar para fora do município de residência para tratamento por quimioterapia e radioterapia é semelhante à de deslocamento para internações hospitalares. Contudo, ao se considerar a frequência com que esses tratamentos necessitam ser realizados no ciclo terapêutico, torna-se preocupante o potencial impacto desse deslocamento na qualidade de vida das mulheres em tratamento. A elevada distância rodoviária percorrida pelos pacientes para internações e tratamento cria dificuldades adicionais ao próprio tratamento e à recuperação pós-cirúrgica.

As estimativas das distâncias rodoviárias percorridas por esses pacientes apontam para uma grande disparidade regional no Brasil. Enquanto os estados das regiões Norte e Nordeste apresentaram maiores deslocamentos rodoviários, principalmente para a Região Sudeste, os estados das regiões Sudeste e Sul apresentam menores deslocamentos necessários para internações e tratamento. Esse comportamento reflete a concentração de unidades de saúde de média e alta complexidade nessas regiões.

Seria possível supor, em princípio, que as distâncias percorridas são maiores em pares de municípios com menor fluxo de pacientes e que as distâncias são menores entre pares de municípios com maior fluxo. Contudo, apesar de o teste de correlação de Spearman entre essas medidas ser significante e negativo nas três redes, seus valores variam entre -0,20 e -0,40. Isso indica que, em boa parte dos pares, a distância entre os municípios não preserva relação com o número de pacientes em deslocamento, conforme ilustrado na Figura 5. Nota-se que um grupo de municípios percorre menores distâncias com grande variação na quantidade de pacientes transportados. Já um outro grupo com menor número de pacientes transportados percorre maiores distâncias.

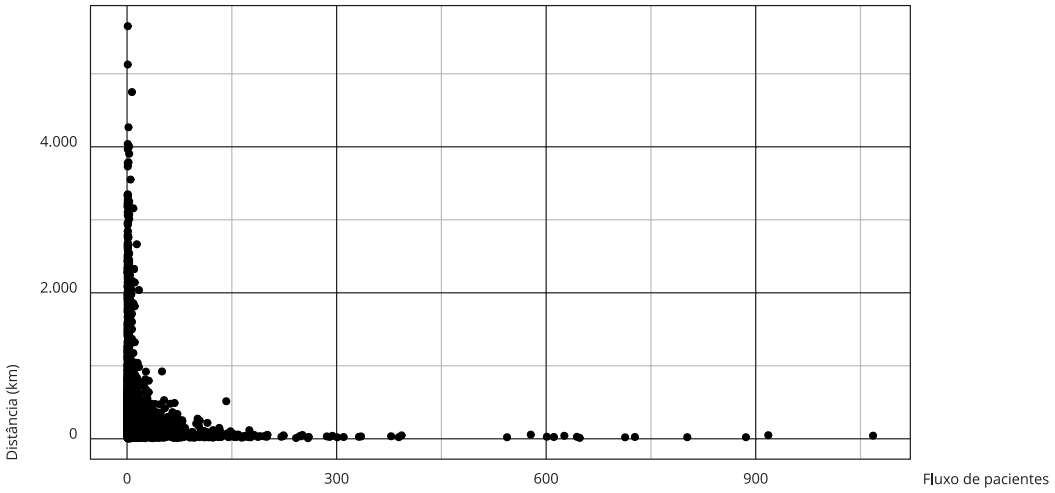
Compreende-se que o deslocamento para acesso a serviços de saúde de média e alta complexidade é esperado em um sistema de saúde hierarquizado. Contudo, os fluxos não previstos e as longas distâncias percorridas pelos pacientes em tratamento evidenciam a necessidade de melhor atuação sobre o planejamento e a regulação dessas redes.

Cabe destacar que a qualidade dos dados sobre local de residência encontrada na AIH e APAC pode apresentar variações, apesar de pesquisa específica sobre a confiabilidade do local de residência para câncer de mama e colo do útero apontar uma confiabilidade de 80% [9].

Como limitação à abrangência deste estudo, os sistemas de informação em saúde utilizados são restritos aos pacientes atendidos no âmbito do SUS, não sendo considerados os pacientes oncológicos atendidos pela rede de saúde suplementar. Essas informações podem tornar ainda mais evidente o

Figura 5

Relação entre fluxo de pacientes (Autorização de Internação Hospitalar – AIH) e distância rodoviária estimada.



baixo deslocamento de pessoas em grandes centros e o maior deslocamento de pacientes nas regiões do Brasil com baixa oferta de atendimento de câncer por planos de saúde.

Percebe-se que alguns fluxos inesperados podem comprometer planos existentes de regionalização da saúde, criando áreas de captação de pacientes muito além dos limites estaduais. Ao preconizar que as regiões de saúde sejam coincidentes geograficamente com as unidades federativas, impõe-se uma necessidade artificial nas centrais de regulação. Isso faz com que pacientes de um estado se desloquem desnecessariamente para sua respectiva capital ou polo regional, embora pudessem ser atendidos em unidades de saúde mais próximas, localizadas em unidades federativas vizinhas.

O estudo de redes específicas para tipos de diagnósticos e regiões, a confrontação entre comunidades detectadas e normatizações existentes de hierarquização e regiões de saúde e, ainda, estudos temporais de redes que evidenciem a conformação e o planejamento das redes podem detalhar e fazer avançar o conhecimento sobre o ordenamento das redes de atendimento à saúde no câncer de mama.

Colaboradores

R. F. Saldanha, D. R. Xavier e C. Barcellos contribuíram com a concepção do projeto, com a redação e com a revisão crítica do artigo. K. M. Carnavalli e K. Lerner contribuíram com a redação e com a revisão crítica do artigo.

Informações adicionais

ORCID: Raphael de Freitas Saldanha (0000-0003-0652-8466); Diego Ricardo Xavier (0000-0001-5259-7732); Keila de Morais Carnavalli (0000-0002-0736-2456); Kátia Lerner (0000-0003-3655-9677); Christovam Barcellos (0000-0002-1161-2753).

Referências

1. World Health Organization. WHO Cancer Fact Sheet 297. http://www.who.int/media-centre/factsheets/fs297/en/ (acessado em 20/Jul/2016).
2. Instituto Nacional de Câncer José Alencar Gomes da Silva. Estimativa 2018: incidência de câncer no Brasil. Rio de Janeiro: Instituto Nacional de Câncer José Alencar Gomes da Silva; 2017.
3. Youlden DR, Cramb SM, Dunn NAM, Muller JM, Pyke CM, Baade PD. The descriptive epidemiology of female breast cancer: an international comparison of screening, incidence, survival and mortality. Cancer Epidemiol 2012; 36:237-48.
4. Ministério da Saúde. Portaria nº 874, de 16 de maio de 2013. Institui a Política Nacional para a Prevenção e Controle do Câncer na Rede de Atenção à Saúde das Pessoas com Doenças Crônicas no âmbito do Sistema Único de Saúde (SUS). Diário Oficial da União 2013; 17 mai.
5. Almeida PF, Giovanella L, Mendonça MHM, Escorel S. Desafios à coordenação dos cuidados em saúde: estratégias de integração entre níveis assistenciais em grandes centros urbanos. Cad Saúde Pública 2010; 26:286-98.
6. Corrêa RL. Interações espaciais. In: Castro IE, Gomes PCC, Corrêa RL, organizadores. Explorações geográficas. Rio de Janeiro: Bertrand Brasil; 1998. p. 279-318. 7. Vian AL, Lima LD, Ferreira MP. Condicionantes estruturais da regionalização na saúde: tipologia dos Colegiados de Gestão Regional. Ciênc Saúde Colet 2010; 15:2317-232.
8. Lima LD, Viana Ald'A, Macgado CV, Albuquerque MV, Oliveira RG, Iozzi FL, et al. Regionalização e acesso à saúde nos estados brasileiros: condicionantes históricos e político-institucionais. Ciênc Saúde Colet 2012; 17:2881-92.
9. Aguiar FP, Melo ECP, Oliveira EXG, Carvalho MS, Pinheiro RS. Confiabilidade da informação sobre município de residência no Sistema de Informações Hospitalares – Sistema Único de Saúde para análise do fluxo de pacientes no atendimento do câncer de mama e do colo do útero. Cad Saúde Colet (Rio J.) 2013; 21:197-200.
10. Grabois MF, Oliveira EXG, Carvalho MS. Assistência ao câncer entre crianças e adolescentes: mapeamento dos fluxos origem-destino no Brasil. Rev Saúde Pública 2013; 47:368-78
11. Oliveira EXG, Melo ECP, Pinheiro RS, Noronha CP, Carvalho MS. Acesso à assistência oncológica: mapeamento dos fluxos origem-destino das internações e dos atendimentos ambulatoriais. O caso do câncer de mama. Cad Saúde Pública 2011; 27:317-26.
12. Souza FS, Silva LMFR, Roveri E. Desenvolvimento de um sistema para o gerenciamento das internações e fluxo de pacientes entre hospitais e cidades de uma região. In: Anais do XI Congresso Brasileiro de Informática em Saúde. Campos do Jordão: Universidade Federal de Minas Gerais; 2008. p. 1-6.

13. Luke DAD, Harris JK. Network analysis in public health: history, methods, and applications. Annu Rev Public Health 2007; 28:69-93.
14. Barabasi A-L. Network science. New York: Cambridge University Press; 2016.
15. Scott J. Social network analysis: a handbook. 2nd Ed. London: Sage Publications; 2000.
16. Wasserman S, Faust K. Social network analysis: methods and applications. Melbourne: Cambridge Press; 1994.
17. Valente TW. Social networks and health: models, methods, and applications. New York: Oxford University Press; 2010.
18. Blanchet K, James P. How to do (or not to do) ... a social network analysis in health systems research. Health Policy Plan 2012; 27:438-46.
19. Sontag S. A doença como metáfora. São Paulo: Edições Graal; 1984.
20. Sousa LMO, Araújo EM, Miranda JGV. Caracterização do acesso à assistência ao parto normal na Bahia, Brasil, a partir da teoria dos grafos. Cad Saúde Pública 2017; 33:e00101616.
21. Borgatti SP, Foster PC. The network paradigm in organizational research: a review and typology. J Manage 2003; 29:99-1013.
22. Ribeiro MGM, Santos SMR, Teixeira MTB. Itinerário terapêutico de mulheres com câncer do colo do útero: uma abordagem focada na prevenção. Rev Bras Cancerol 2011; 57:48-91.
23. Azevedo e Silva G, Bustamante-Teixeira MT, Aquino EML, Tomazelli JG, Dos-Santos-Silva I. Acesso à detecção precoce do câncer de mama no Sistema Único de Saúde: uma análise a partir dos dados do Sistema de Informações em Saúde. Cad Saúde Pública 2014; 30:1537-50.
24. Mancini DVG. Fluxo da assistência oncológica em Minas Gerais a partir das informações sobre os óbitos por câncer de mama em mulheres [Dissertação de Mestrado]. Juiz de Fora: Universidade Federal de Juiz de Fora; 2015.

Abstract

This study aims to analyze the flow of breast cancer patients treated outside of their municipality of residence, based on hospital admissions and chemotherapy and radiotherapy in the Brazilian Unified National Health System (SUS) from 2014 to 2016. Network analysis was used, considering the municipality of residence and of treatment as nodes in a graph, thus consisting of a "health system organizational network study". In addition, highway distances and travel time were estimated via the best feasible route according to the Open Street Maps *highway project. According to the results, 51.34% of breast cancer patients in Brazil were treated outside their municipality of residence, following regionalized flows that respect state borders, generally towards the state capital or other large cities. The results also point to specific exceptions, where some municipalities occupy outstanding positions that extrapolate state borders. Median travel time from the municipality of residence to the municipality of care was nearly 3 hours, and 75% of trips totaled 324km for chemotherapy, 287km for radiotherapy, and 282km for hospitalizations. These results are indicative of the difficulties in access to oncology services, potentially aggravating the illness experience with cancer in terms of impact on the individuals and their families.*

Breast Neoplasms; Health Services Accessibility; Information Systems

Resumen

El objetivo de este estudio fue analizar el flujo de pacientes oncológicos con cáncer de mama que son atendidos fuera de su domicilio de residencia. Se consideraron internamientos hospitalarios, tratamientos por quimioterapia y radioterapia para neoplasias malignas de mama, dentro del ámbito del Sistema Único de Salud brasileño, entre los años de 2014 a 2016. Se empleó el método de análisis de redes, considerando como nudos de un grafo el municipio de residencia y el del tratamiento, formándose de esta forma un "estudio de redes organizativas de sistemas de salud". Asimismo, se estimaron las distancias viales y el tiempo de desplazamiento, a través de la mejor ruta de carreteras, según la red de carreteras del proyecto Open Street Maps*. Los resultados apuntan que un 51,34% de los pacientes con cáncer de mama en Brasil fueron atendidos fuera de su municipio de residencia, siguiendo flujos regionalizados y dentro de sus fronteras estatales, en general, en dirección a las capitales de las mismas o grandes ciudades. Por otro lado, los resultados también muestran excepciones específicas, donde algunos municipios detentan un grado de relevancia superando las fronteras estatales. El tiempo de desplazamiento entre el municipio de residencia y el municipio de atención presentó unas medias cercanas a las 3 horas, y en un 75% de los desplazamientos se recorrieron hasta 324km para recibir tratamiento de quimioterapia, 287km para el tratamiento de radioterapia y 282km para internamientos. Estos resultados son indicativos de las dificultades de acceso a los servicios de oncología, lo que agrava potencialmente la experiencia de la enfermedad oncológica en términos de impacto en el individuo y su familia.*

Neoplasias de la Mama; Accesibilidad a los Servicios de Salud; Sistemas de Información

Recebido em 08/Mai/2018
Versão final reapresentada em 18/Fev/2019
Aprovado em 21/Fev/2019

4.4 Da coleta distribuída à visualização

O artigo apresentado abaixo, intitulado "Contributing to elimination of cross-border malaria through a standardized solution for case surveillance, data sharing, and data interpretation: development of a cross-border monitoring system" é resultado de uma parceria internacional denominada *Laboratório Misto Internacion* (LMI), coordenada por pesquisadores da Fiocruz, Universidade de Brasília (UnB) e o Instituto francês de Pesquisa para o Desenvolvimento (IRD).

Neste artigo, um estudo de caso completo da aplicação do modelo de processos KDD é apresentado, iniciando pela aquisição de dados de diversas fontes e diferentes formatos, passando por etapas de pré-processamento e harmonização de dados até a apresentação dos resultados através de visualizações interativas de dados na Internet.

O artigo foi publicado no *Journal of Medical Internet Research – Public Health and Surveillance* em setembro de 2020.

Original Paper

Contributing to Elimination of Cross-Border Malaria Through a Standardized Solution for Case Surveillance, Data Sharing, and Data Interpretation: Development of a Cross-Border Monitoring System

Raphael Saldanha[1,2], MSc; Émilie Mosnier[3,4], MD, PhD; Christovam Barcellos[1,2], PhD; Aurel Carbunar[3], MSc; Christophe Charron[2,5], MSc; Jean-Christophe Desconnets[2,5], PhD; Basma Guarmit[3], MSc; Margarete Do Socorro Mendonça Gomes[6], PhD; Théophile Mandon[5], MSc; Anapaula Martins Mendes[7], MSc; Paulo César Peiter[2,8], PhD; Lise Musset[9,10], PharmD, PhD; Alice Sanna[11], MSc; Benoît Van Gastel[11], MSc; Emmanuel Roux[1,2,5], PhD

[1]Laboratório de Informação em Saúde, Instituto de Comunicação e Informação Científica e Tecnológica em Saúde, Fundação Oswaldo Cruz, Rio de Janeiro, Brazil

[2]Laboratoire Mixte International Sentinela, Fundação Oswaldo Cruz, Universidade de Brasília, Institut de Recherche pour le Développement, Rio de Janeiro, Brazil

[3]Service des Centres Délocalisés de Prévention et de Soins, Centre Hospitalier de Cayenne, Cayenne, French Guiana

[4]Sciences Économiques et Sociales de la Santé et Traitement de l'Information Médicale, Aix Marseille Université, Institut National de la Santé et de la Recherche Médicale, Institut de Recherche pour le Développement, Marseille, France

[5]Espace-Dev, Institut de Recherche pour le Développement, Université de Montpellier, Université de La Réunion, Université de Guyane, Université des Antilles, Cayenne, French Guiana, and Montpellier, France

[6]Superintendência de Vigilância em Saúde do Estado do Amapá, Macapá, Brazil

[7]Universidade Federal do Amapá, Oiapoque, Brazil

[8]Laboratório de Doenças Parasitárias, Instituto Oswaldo Cruz, Fundação Oswaldo Cruz, Rio de Janeiro, Brazil

[9]Laboratoire de Parasitologie, Institut Pasteur de la Guyane, Cayenne, French Guiana

[10]Centre National de Référence du Paludisme, Pôle Zones Endémiques Françaises, World Health Organization Collaborating Center for Surveillance of Antimalarial Drug Resistance, Cayenne, French Guiana

[11]Agence Régionale de Santé de Guyane, Cayenne, French Guiana

Corresponding Author:
Christovam Barcellos, PhD
Laboratório de Informação em Saúde
Instituto de Comunicação e Informação Científica e Tecnológica em Saúde
Fundação Oswaldo Cruz
Avenida Brasil, 4365
Manguinhos
Rio de Janeiro, 21040-360
Brazil
Phone: 55 2138653242
Email: christovam.barcellos@fiocruz.br

Abstract

Background: Cross-border malaria is a significant obstacle to achieving malaria control and elimination worldwide.

Objective: This study aimed to build a cross-border surveillance system that can make comparable and qualified data available to all parties involved in malaria control between French Guiana and Brazil.

Methods: Data reconciliation rules based on expert knowledge were defined and applied to the heterogeneous data provided by the existing malaria surveillance systems of both countries. Visualization dashboards were designed to facilitate progressive data exploration, analysis, and interpretation. Dedicated advanced open source and robust software solutions were chosen to facilitate solution sharing and reuse.

Results: A database gathering the harmonized data on cross-border malaria epidemiology is updated monthly with new individual malaria cases from both countries. Online dashboards permit a progressive and user-friendly visualization of raw data and epidemiological indicators, in the form of time series, maps, and data quality indexes. The monitoring system was shown to be able to identify changes in time series that are related to control actions, as well as differentiated changes according to space and to population subgroups.

Conclusions: This cross-border monitoring tool could help produce new scientific evidence on cross-border malaria dynamics, implementing cross-border cooperation for malaria control and elimination, and can be quickly adapted to other cross-border contexts.

(JMIR Public Health Surveill 2020;6(3):e15409) doi: 10.2196/15409

KEYWORDS

cross-border malaria; surveillance; data interoperability; data visualization; French Guiana; Brazil

Introduction

The Global Technical Strategy of the World Health Organization (WHO) [1] aims for a 90% reduction in global malaria mortality and incidence by 2030 in comparison with 2015 levels, notably by "transforming malaria surveillance into a core intervention."

However, several obstacles make such a strategy difficult to apply and the elimination target challenging to reach. One of them is *cross-border malaria* [2-7]. Cross-border malaria does not only refer to the malaria cases that cross international borders, but also to all aspects of the disease within cross-border living territories that require actual cross-border visions. However, from one country to another, differences are observed in disease diagnosis and treatment protocols, the epidemiological information collected, database structures, information representations (ie, database attribute names, formats, encoding, etc), data access protocols and rights, and so forth. Such differences prevent the border countries from having a shared and unified view of the cross-border epidemiological situation and, thus, to jointly design and implement efficient control actions. Cross-border epidemiological surveillance systems are required to overcome such obstacles. One solution is to build them into existing national systems, when they exist, by ensuring data interoperability. However, data reconciliation implies dealing with semantic, structural, and syntactic heterogeneities. Moreover, the diversity of recipients of the harmonized data (ie, health actors, health and territory managers, the general public, etc) challenges the actual and advantageous dissemination of cross-border harmonized data and knowledge. In fact, the potential recipients differ notably in their objectives, background knowledge on the disease, technological skills, and languages.

The French Guiana–Brazil border is an endemic malaria region [8]. The Franco-Brazilian cooperation agreement of May 28, 1996, led to the creation of the Joint Commission for Cross-Border Cooperation between French Guiana and Brazil. A subworking group has been working exclusively on health-related issues since 2009. Notably, this resulted in regular epidemiological data exchanges on malaria between French Guianese and Brazilian malaria surveillance authorities. However, differences in data formats, update frequencies, spatial and temporal aggregation units, and nature of information; the lack of contextual information (ie, metadata) and shared frame of reference, notably, a cartographic representation; as well as the limited numbers of recipients of the information on both sides of the border make such a procedure inefficient in providing a unified vision of the malaria situation in the cross-border area. This consequently prevents the design and implementation of concerted control and elimination actions.

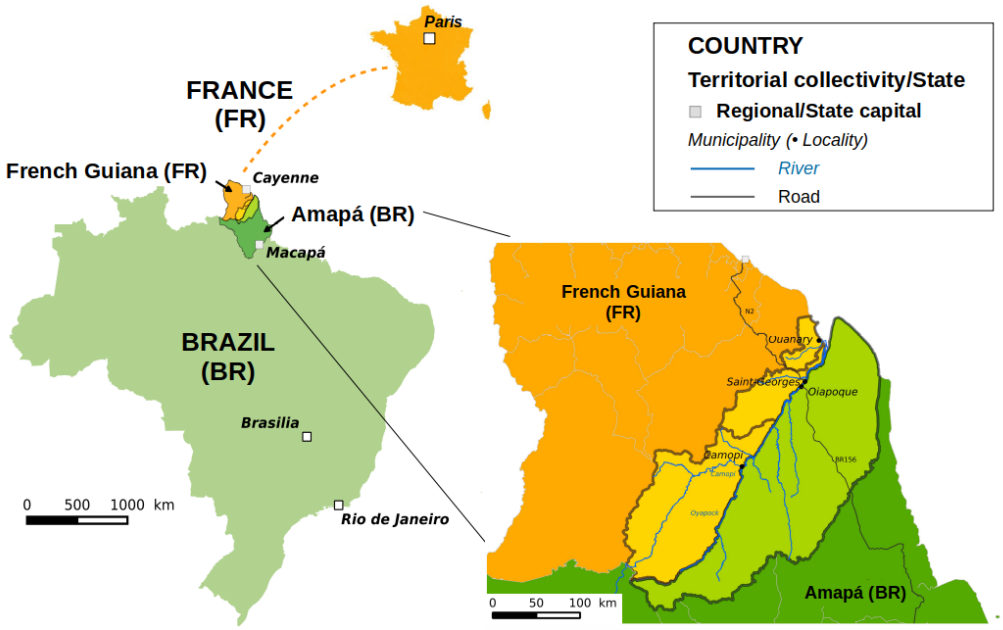
In this context, building a cross-border malaria information system (CBMIS) is needed. This requires specifying easily reproducible methods based on explicit data harmonization rules, free technological solutions, as well as information representation and dissemination good practices. Moreover, data visualization solutions for health actors, health and territory managers, and the general public are necessary to facilitate data and knowledge dissemination. This paper addresses such issues by describing a cross-border system for data harmonization and visualization implemented between French Guiana and Brazil.

Methods

Study Area

French Guiana—83,534 km^2 in area with an estimated 290,691 inhabitants in 2020 [9]—is a French overseas region located in the Amazon, South America. French Guiana consists of 22 municipalities, with four of them bordering Brazil: Maripasoula, Camopi, Saint-Georges de l'Oyapock (hereafter referred to as Saint-Georges), and Ouanary. Amapá—142,829 km^2 in area with an estimated 845,731 inhabitants in 2019 [10]—is one of the 27 states, including the federal district, of the Federative Republic of Brazil. The Amapá state is located in the Brazilian Amazon, bordering French Guiana to the north (see Figure 1).

Figure 1. Cross-border area delimitation and administrative structuration of the region.



For the development of the CBMIS, the cross-border area between French Guiana and Brazil was defined by the border municipalities of both countries, which define a coherent and continuous living territory for local populations (see Figure 1): for French Guiana, this includes Ouanary, Saint-Georges, and Camopi, with 201, 4220, and 1828 inhabitants in 2017, respectively [9]; for Brazil, this includes Oiapoque, with 27,270 inhabitants in 2019 [10]. The population living in this area is distributed over two main urban centers, Saint-Georges and Oiapoque, as well as in villages mainly located along the Oiapoque River, along the BR-156 road in Amapá, and in territories with restricted access (ie, natural parks on both sides of the border and the Brazilian Amerindian Territories).

Data Sources and Definition of Cross-Border Malaria Cases

Concerning French Guiana, anonymized information regarding *individual malaria cases* is collected monthly from the surveillance system of the delocalized Centers for Prevention and Care (Centres Délocalisés de Prévention et de Soins [CDPSs]) operated by the Cayenne Hospital, which has been operating since 2007. Four CDPSs are present in the cross-border area: in Ouanary, Saint-Georges, Camopi, and Trois-Sauts (Camopi municipality). In this system, a malaria case is defined as any positive rapid diagnostic test (RDT) (SD BIOLINE Malaria Ag Pf/Pan in French Guiana). Such tests only distinguish *P falciparum* and non-*P falciparum* species. *New attacks* of malaria (ie, new infections due to new mosquito bites, to be distinguished from malaria notifications related to the follow-up of patients, treatment failures, or *P vivax* relapses) are not explicitly identified in the database. Each patient in the database is identified by a unique coded identifier.

Regarding Brazil, information on *individual malaria cases* is provided by the Malaria Epidemiological Surveillance Information System (Sistema de Informações de Vigilância Epidemiológica da Malaria [SIVEP-Malária]), operated by the information technology department of the unified health system (Departamento de Informática do Sistema Único de Saúde) of the Brazilian Ministry of Health. Brazil mainly uses thick smear microscopy, allowing for the identification of all *Plasmodium* species and development stages, but also the RDT (SD BIOLINE Ag Pf/Pf/Pv).

In the Brazilian database, malaria attacks related to follow-up consultations, treatment failures, and relapses are all referred to as *treatment verification slides* (lâminas de verificação de cura [LVCs]). A malaria case is considered as an LVC for *P vivax* (or for *P falciparum*) if the patient received treatment against *P vivax* (or for *P falciparum*) within the last 60 days (40 days for *P falciparum*) [11]. A non-LVC case is considered a *new case*. Patients are not identified by a unique coded identifier. The SIVEP-Malária supplies anonymized data on a monthly basis to the CBMIS through a partnership with the Oswaldo Cruz Foundation (Fundação Oswaldo Cruz [Fiocruz]). Database fields of the French and Brazilian surveillance systems that were considered in the CBMIS are detailed in Multimedia Appendix 1, Table S1.

A *cross-border malaria case* was defined as any malaria case as defined by the national surveillance systems and that was associated with (1) a notification center, (2) a patient's residential address, *or* (3) a possible transmission location, *located in the previously defined cross-border area*.

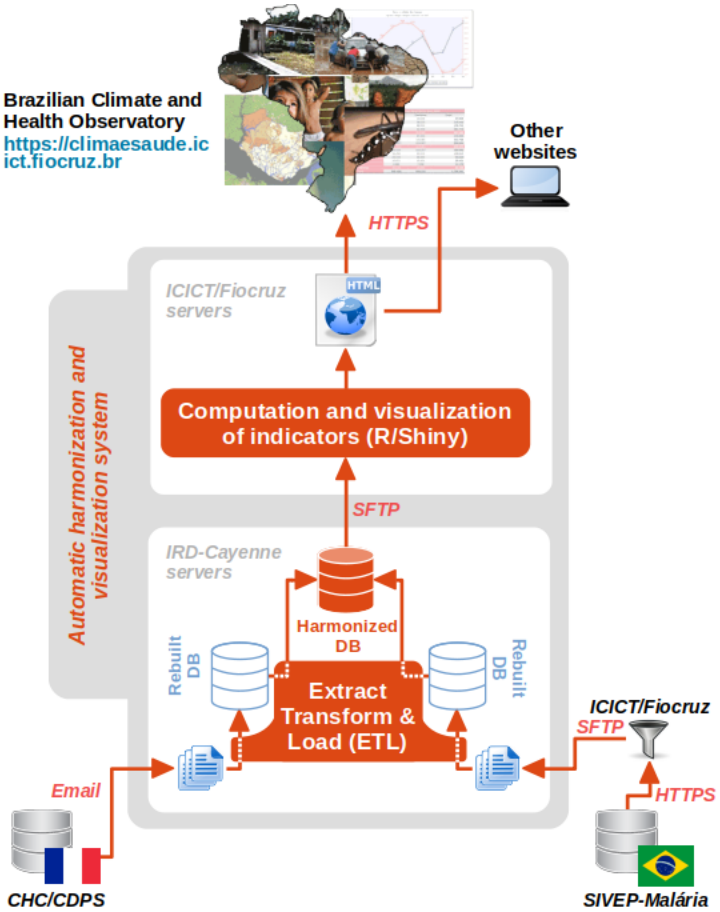
The two surveillance systems report on the locations of notification centers, residences, or putative contamination locations, with respect to predefined and scalable lists of *localities* (ie, a locality being either isolated but inhabited places, villages, or urban neighborhoods), but without systematically providing their geographical coordinates [12]. Thus, geographical coordinates of localities were obtained through various sources: knowledge of the researchers and partners

involved in the project; OpenStreetMap collaborative project; National Indigenous Foundation (for Brazilian Amerindian villages); Google and Bing satellite imagery; and Sentinel-2 satellite images from the European Space Agency, retrieved from the operating platform (Plateforme d'Exploitation des Produits Sentinel) of the Sentinel products developed by the French space agency (Centre National d'Études Spatiales).

Data Harmonization System

Harmonization was aimed at transforming the data from the two national information systems in order to make them satisfy a common harmonized data model; see Figure 2 for a representation of the global data flow, with the main harmonization steps and the data transfer protocols used.

Figure 2. Overall system architecture and data and information flow. CDPS: Service des Centres Délocalisés de Prévention et de Soins (Department of the Centers for Prevention and Care); CHC: Centre Hospitalier de Cayenne (Cayenne Hospital); DB: database; Fiocruz: Fundação Oswaldo Cruz (Oswaldo Cruz Foundation); HTTPS: hypertext transfer protocol secure; ICICT: Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (Institute of Scientific and Technological Communication and Information in Health); IRD: Institut de Recherche pour le Développement (French National Research Institute for Sustainable Development); SFTP: secure shell file transfer protocol; SIVEP-Malária: Sistema de Informações de Vigilância Epidemiológica da Malária (Malaria Epidemiological Surveillance Information System).



This common harmonized data model relied, as much as possible, on existing standards: international standards or, if not available, national ones or even de facto normative representations, due to their extensive and consensual use in the knowledge areas involved in the study. In practice, harmonization consisted of changes in data types (eg, conversion from string type to integer type for the sex field in the SIVEP-Malária database), unit conversions (eg, patient age conversion from days or months to years), and data transformations that required more deep knowledge on malaria surveillance and parasitology, especially regarding *Plasmodium* species specification and new malaria case detection. The information provided by the RDT on *Plasmodium* species was more general and was the only information shared by both countries. In the harmonized database, *Plasmodium* species were consequently coded as "*P falciparum*," "non-*P falciparum*," "mixed infection with *P falciparum*," or "Unspecified" (see Multimedia Appendix 1, Table S2, for details). Eventually, a *new attack* was defined in the CBMIS: for data from the SIVEP-Malária (Brazil), this was defined as any case notification that is *not* an LVC; for data from the CDPS database (French Guiana), this was defined as any *P vivax* (or *P falciparum*) case notification that occurs at least 91 days (41 days for *P falciparum*) after the last *new attack* of *P vivax* (or *P falciparum*). In fact, French epidemiologists consider that a *P vivax* malaria notification can be considered as a *new case* if it occurs more than 90 days after the last contamination [13].

Unique patient identifiers were used to reconstruct the patient notification history and to apply this *new case* detection rule.

The initial data representations within the national systems, the harmonized data model, and associated standards, as well as the harmonization rules, are provided in Multimedia Appendix 1, Table S1.

An *extract, transform, and load* (ETL) process, implemented by the free software Talend Open Studio for Big Data, was used to apply all the transformation rules.

Harmonized Data Visualization and Dissemination

To deal with the previously mentioned barriers to information and knowledge dissemination, progressive access to information was implemented using the Shneiderman et al mantra [14]: "Overview first, zoom and filter, then details-on-demand." Dashboards in three languages—Portuguese, French, and English—accessible to the users via the internet, using any updated browser on a computer or mobile device, were developed. The visualization tool has been implemented in two versions: a *general public* version, accessible without any authentication procedure but with restricted functionalities and data access, and an *expert* version, accessible through log-in and password and with full access to master harmonized data and functionalities. Multimedia Appendix 1, Table S3, details the functionalities of the two versions.

The visualization dashboards were implemented with the R package Shiny (RStudio) [15]. They were made accessible online [16,17]. Access to dashboards was also provided through the Brazilian Climate and Health Observatory [18], more precisely via the webpage dedicated to the Amapá–French Guiana *surveillance area* [19].

Legal and Ethical Considerations

Data on malaria cases are received already anonymized from the CDPS department and the SIVEP-Malária. The CBMIS ensures the automatic processing of patient-related personal data and the transfer of these data to the Brazilian partner. This required the following: (1) the authorization from the French data protection authority (Commission Nationale de l'Informatique et des Libertés [CNIL]), which verifies compliance with the General Data Protection Regulation (EU) 2016/679 (CNIL deliberation No. 2019-025 of 28 February 2019, request for authorization No. 2135363), and (2) the ratification of the *European Union standard contractual clauses for transfers between two data controllers*. In Brazil, all the actions carried out were authorized as part of the Fiocruz public health activities, as per the Brazilian *free access* law 12.527 of November 18, 2011, and in compliance with law 13.709 of August 14, 2018.

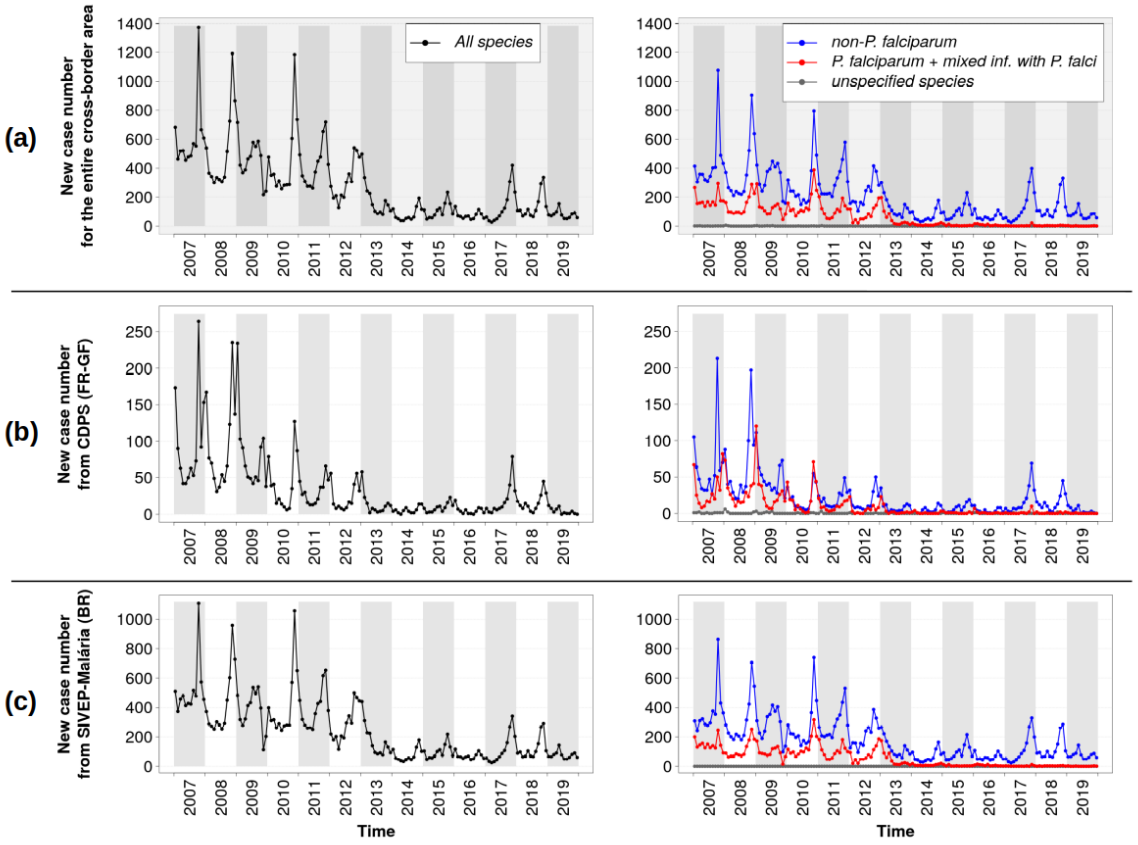
The compliance with legal requirements demanded a specific algorithmic development for new case identification in the French Guiana database, which is detailed in Multimedia Appendix 1, Figure S1.

Results

The CBMIS has been implemented and updated and harmonized data are delivered monthly. Data are available starting from 2003 and 2007 for the SIVEP-Malária Brazilian system and the CDPS French Guiana database, respectively. Some key harmonized database contents for the common period (ie, since 2007) are presented hereafter.

Figure 3 shows the number of new malaria cases in the cross-border area as a whole and as a function of the country of notification.

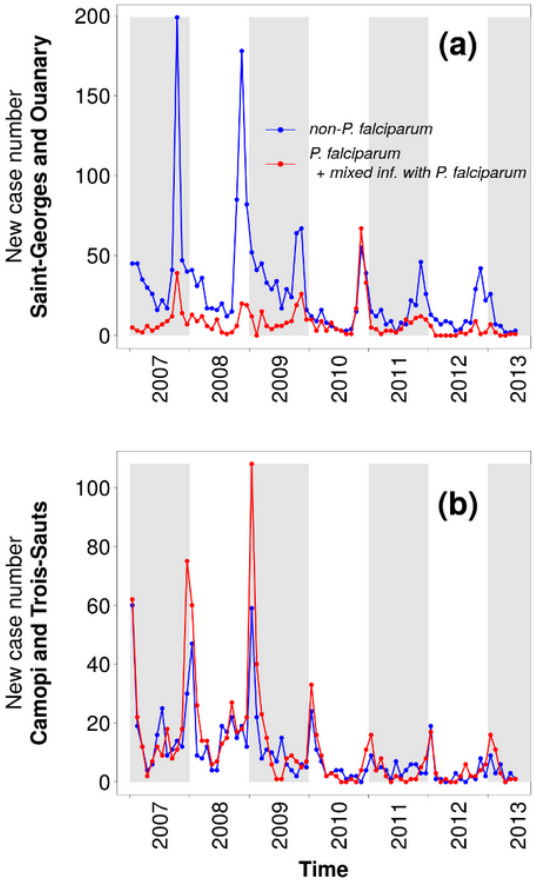
Figure 3. Number of new monthly malaria cases reported in the cross-border area from 2007 to 2019: (a) the cross-border area as a whole; (b) cases recorded in the database of the Department of the Centers for Prevention and Care (Service des Centres Délocalisés de Prévention et de Soins [CDPS]) in French Guiana (FR-GF); (c) cases recorded in the Malaria Epidemiological Surveillance Information System (Sistema de Informações de Vigilância Epidemiológica da Malária [SIVEP-Malária]) in Brazil (BR).



Cases notified by both countries, globally, presented comparable dynamics, with a clear seasonality showing a peak between October and December (ie, at the end of the dry season and the early beginning of the rainy season). Four main phases can be distinguished over the total period:

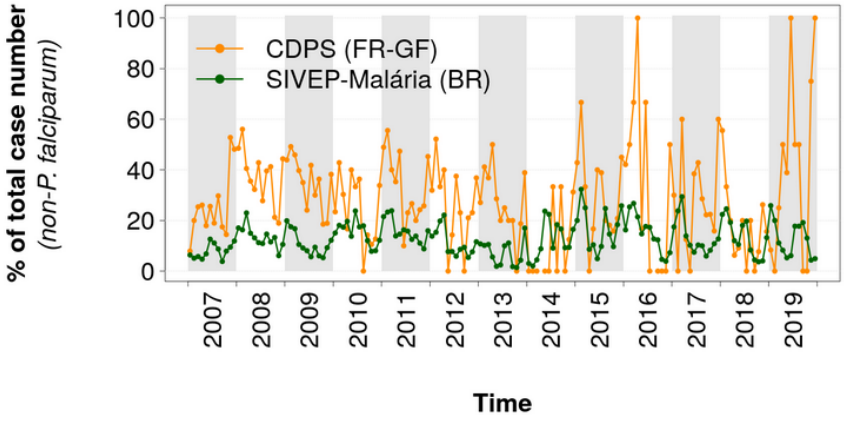
1. January 2007 to June 2013: high but decreasing number of cases. Figure 3 (b) shows a two-peak epidemic curve in cases notified in the CDPS database (French Guiana) for this period, except for the year 2010. These two peaks were associated with different subregions and, to a lesser extent, with different *Plasmodium* species (see Figure 4). The first peak (October to November) corresponded with the lower Oyapock River region (ie, Saint-Georges and Ouanary), with a majority of non-*P falciparum* cases, as seen in Figure 4 (a); the second peak (December to January) corresponded to the upper Oyapock River region (ie, Trois Sauts and Camopi), with a majority of *P falciparum* cases, as seen in Figure 4 (b). Moreover, two subphases can be seen during this period in the cases provided by the CDPS database: a high and quite stable number of cases in 2007 and 2008 and a significant drop in the number of cases in 2009, followed by a progressive decrease up to 2013.
2. July 2013 to December 2016: low number of cases with relative interannual stability, despite a higher number of cases in 2015. The year 2013 particularly corresponded to a significant drop in the number of *P falciparum* cases (see Figure 3).
3. January 2017 to December 2018: recrudescence of *P vivax* cases.
4. January 2019: number of cases comparable with the 2013-2016 period, even lower for CDPS data, with a peak earlier in the year in May, particularly marked in the data provided by the SIVEP-Malária (Brazil).

Figure 4. Monthly reported malaria cases by species at the Centers for Prevention and Care (Centres Délocalisés de Prévention et de Soins [CDPSs]) of (a) Saint Georges de l'Oyapock and Ouanary and (b) Camopi and Trois Sauts, between January 2007 to June 2013.



For non-*P falciparum* species, a significantly higher percentage of cases related to follow-up, treatment failures, and relapses were identified in the CDPS database (see Figure 5). During the whole period, the average percentages were 28.7% and 12.7% in the CDPS database and in the SIVEP-Malária, respectively. As the number of cases became very low in French Guiana in 2016 and 2019, no malaria case was reported for some months; for other months, 100% of the cases were associated with follow-ups, putative treatment failures, or relapses.

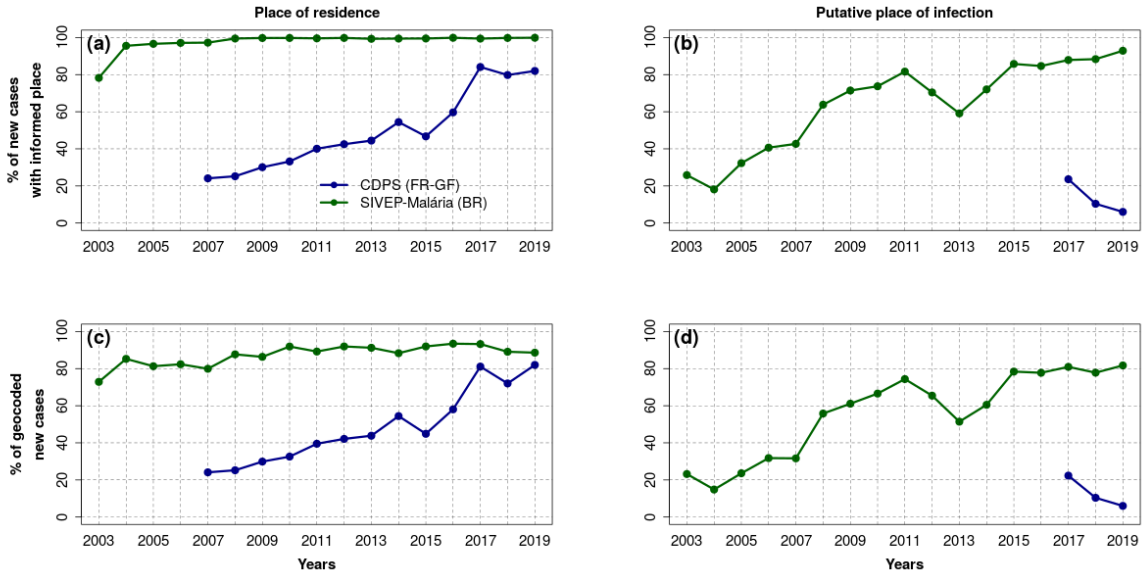
Figure 5. Percentages of cases associated with follow-up, treatment failures, or relapses for non-P falciparum cases in the database of the Department of the Centers for Prevention and Care (Service des Centres Délocalisés de Prévention et de Soins [CDPS]) in French Guiana (FR-GF) and the Malaria Epidemiological Surveillance Information System (Sistema de Informações de Vigilância Epidemiológica da Malária [SIVEP-Malária]) in Brazil (BR).



In the CDPS database, the percentage of cases associated with a place of residence increased from less than 30% in 2007 to more than 80% since 2017, as seen in Figure 6 (a). On the other hand, 100% of the new cases from the SIVEP-Malária database

were associated with a place of residence since 2008, as seen in Figure 6 (a).

Figure 6. Percentage of malaria cases in the database of the Department of the Centers for Prevention and Care (Service des Centres Délocalisés de Prévention et de Soins [CDPS]) in French Guiana (FR-GF) and in the Malaria Epidemiological Surveillance Information System (Sistema de Informações de Vigilância Epidemiológica da Malária [SIVEP-Malária]) in Brazil (BR) associated with (a) a place of residence; (b) a putative place of infection; (c) a geolocalized place of residence; and (d) a geolocalized putative place of infection. Putative places of infection were not stored in the CDPS database before 2017.

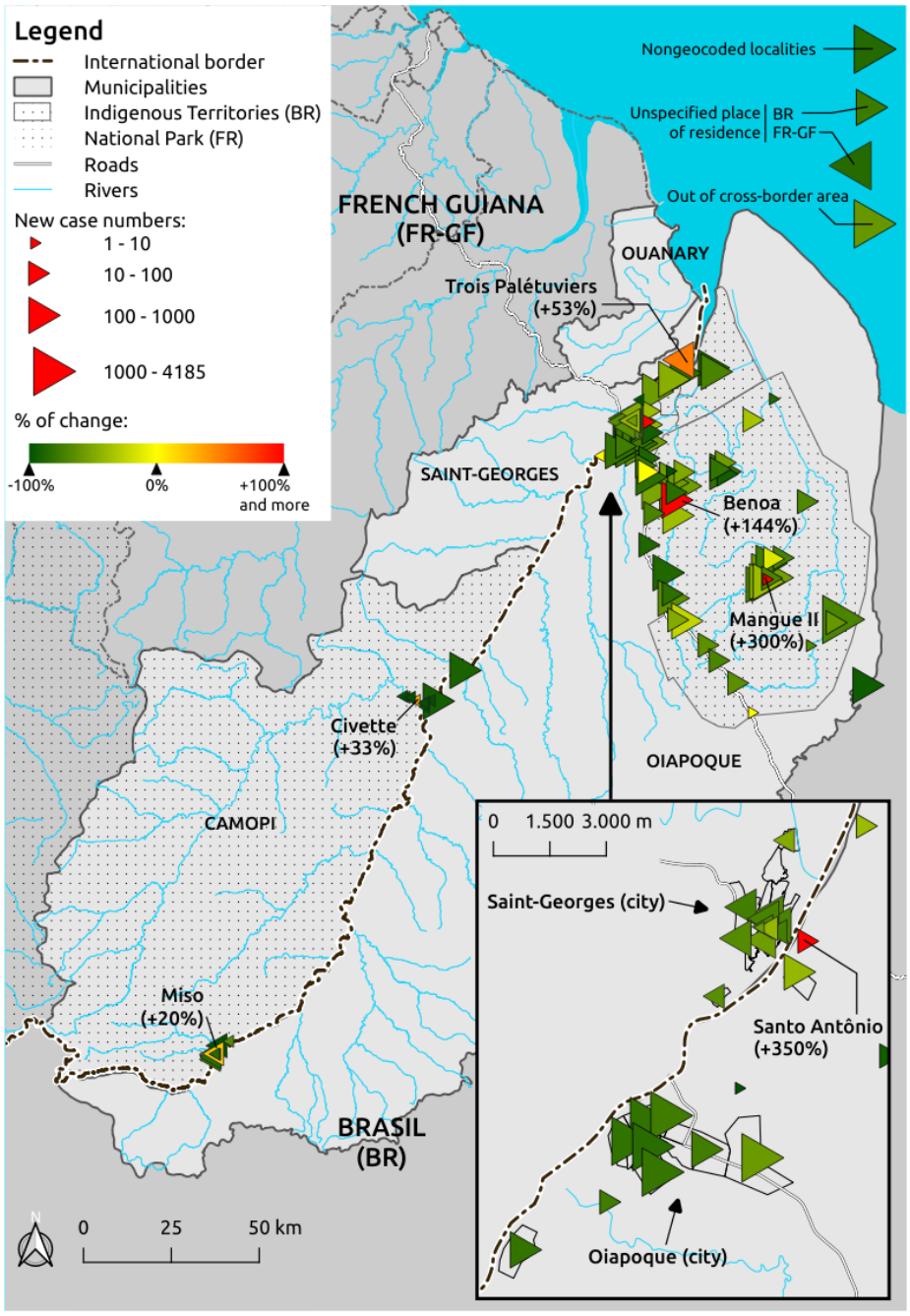


Concerning the putative place of infection of the new cases, the information has only been stored in the CDPS database since 2017. Such information remained rare and even tended to be rarer in the CDPS database, passing from about 20% of the new cases in 2017 to less than 10% in 2019 as seen in Figure 6 (b). In the SIVEP-Malária database, such information was much more present, with more than 80% of the new cases associated with a possible place of infection since 2015 as seen in Figure 6 (b).

The specific work carried out in this study to geolocalize, or geocode, localities resulted in 100% and 52.4% of geolocalized localities of the cross-border area for the French Guiana and Brazilian sides, respectively. However, in the SIVEP-Malária, the relatively small proportion of geolocalized localities (52.4%) had little impact on the number of cases actually geolocalized, with about 90% and 80% of the cases geocoded since 2015 in relation with the places of residence and probable places of infection, respectively, as seen in Figure 6 (c) and (d).

Figure 7 shows an example of a map realized with the harmonized data of the CBMIS. It represents the numbers of new cases as a function of the places of residence of the patients, from January 2007 to December 2019, jointly with the percentage of change in the case numbers between the two main periods previously described: January 2007 to June 2013 and July 2013 to December 2019. The map shows a significant decrease in almost the entire cross-border area. The decrease was very significant in the Camopi municipality and in the urban quarters of the Oiapoque city. The decrease was significant but less important in the Amerindian communities of the Oiapoque municipality and in the Saint-Georges municipality. Some localities experienced an increase in case numbers between the two periods: the two Amerindian localities Benoa and Trois Palétuviers, in Brazil and French Guiana, respectively, had a significant increase from 34 to 84 cases (144%) and 93 to 142 cases (53%), respectively. The Amerindian locality Mangue II (Brazil), the locality Santo Antônio (Brazil), as well as the two Amerindian localities Civette and Miso in the Camopi municipality (French Guiana) experienced a nonsignificant increase in regard to the total number of cases.

Figure 7. Number of reported malaria cases as a function of patients' places of residence. Triangles with apexes oriented to the right correspond to Brazilian localities; triangles with apexes oriented to the left correspond to French localities. The triangle size is a function of the case number. The triangle color is a function of the percentage of change in the case number between the following two periods: January 2007 to June 2013 and July 2013 to December 2019.



Discussion

Principal Findings

The results showed the potential of the CBMIS for the analysis of cross-border malaria dynamics, in both space and time. Such a system also allows for pointing out similarities and differences in the epidemiological situations of both countries. As it is shown hereafter, such similarities and differences can be interpreted in terms of control strategies. In the following paragraphs, methodological aspects of the proposed approach and the previously presented results are discussed. However, specific and deep investigations of cross-border epidemiological issues are out of the scope of this paper.

Definition of Cross-Border Malaria Cases

Human mobility is an important issue when considering border regions [2]. By differentiating between places of residence, notification, and infection, the CBMIS allows an estimation of internal and external flows in the area and facilitates the identification of autochthonous and imported malaria cases. Such differentiation also allows for conducting studies from different viewpoints, notably on environmental determinants of the transmission, population profiles, identification of spatial clusters of malaria cases, provision of and access to care, and activity level of health infrastructures.

General Harmonization Strategy

The chosen approach relies on current national health system data reconciliation and does not require any previous system modifications. Such an approach is comparable to the one in Dell'Erba et al [20], which was developed for the domains of travel and tourism information systems and data, or Zinszer et al [21] for malaria data integration. This approach is likely to facilitate the participation of surveillance agencies in the development of a CBMIS, whereas these agencies would be "reluctant to abandon their own data schemata in favor of a standard schema supplied by someone else" [20]. In that sense, the proposed approach differs from recommendations provided in D'Agostino et al [22] to facilitate data sharing in public health, which include the development of regional frameworks that "can be adopted or adapted by each country through national or subnational policies" as a prerequisite for the realization of data interoperability.

In Al Manir et al [23], the authors developed a set of services to query multisource heterogeneous malaria-related data using standard terminologies and rules to match database fields and controlled vocabularies. They illustrated the functioning of the system by answering thematic questions provided by the Uganda Ministry of Health and by querying two data repositories: the Scalable Data Integration for Disease Surveillance platform [21] and the Global Malaria Mapper from the WHO, now integrated into the Global Health Observatory data [24]. The system was not designed to provide and visualize comparable and qualified raw epidemiological data as in this study. However, it can automatically identify any change in source databases and provides tools to reconfigure the system in order to maintain its integrity, unlike our method. Such functionality would be of interest in applying the approach proposed in this article to a large number of surveillance systems.

Data Completeness, Quality, and Limitations

In French Guiana, CDPSs are not the only malaria notifiers. Nevertheless, given the care pathway of the people living in or frequenting the three border municipalities, the quasi-totality of the malaria cases is retrieved by the system. On the other hand, the three French Guiana border municipalities have only been reporting putative places of infection since 2017, and a lot of missing data are associated with this field. As a consequence, some malaria cases can be omitted by the system if their notifications and places of residence are out of the cross-border area, but the putative places of infection would belong to it. However, we can expect such a number to be negligible. In Brazil, the legal Amazon, whose malaria cases are reported in the SIVEP-Malária, accounts for more than 99% of the Brazilian malaria cases [25,26]. In conclusion, the CBMIS reports reliably on the number of cases within the cross-border area.

Some database attributes exhibit a lot of missing data. Among them, the putative place of contamination, and to a lesser extent the place of residence, is by far the least informed in the CDPS database. However, the information on putative places of contamination has been collected for a long time in French Guiana and has been used for malaria control. The epidemiological bulletins on malaria in French Guiana, published by the national agency for epidemiological surveillance (Santé Publique France), reported that, for the whole French Guiana area and the period between January 2017 and September 2019, the suspected place of contamination is known for 76.9% of cases on average, with a global upward trend (minimum of 54.4% for the first trimester of 2017; maximum of 87% for the first trimester of 2019) (see Multimedia Appendix 1, Table S4). These numbers are comparable with those on the Brazilian side and considerably contrast with those previously shown for French Guiana. In fact, when the CDPS transmits the information on new malaria cases to the local health surveillance authority, the latter requests that the vector control service of the French Guiana territorial collectivity carry out intradomiciliary insecticide spraying and to investigate the context of contamination, in particular, the putative place of contamination. There is currently no back-feeding of the CDPS database with the collected information, which should be considered in the future.

It is worth noting that, despite the difficulties encountered in geocoding all localities on the Brazilian side, the great majority of the new cases reported in Brazil are finally geocoded according to their residence and the place of infection. In fact, only very small localities, and localities that no longer exist, that are associated with very low numbers of cases could not be geocoded. However, efforts are continuing to reach the target of 100% geocoded localities on the Brazilian side.

Some of the missing information in the harmonized database may be due to inadequate coding of the information at the time of notification. However, all possible errors cannot be anticipated and considered within an automatic processing framework unless a highly specific system is built, the functioning of which may become difficult to understand and maintain. The strategy chosen for the CBMIS is instead to provide quality indicators, especially relative to missing information, in order to (1) provide users with the primary interpretation keys in order to let them decide whether an information item is significant or not and (2) give feedback to health actors in charge of surveillance, to allow them to identify surveillance system weaknesses and improve their practice.

The far more difficult point is the interpretation biases derived from differences in country surveillance cultures and practices. Some of these differences are not surmountable, and the harmonization requires making choices and compromises, as with the *new attack* notion discussed above and in Multimedia Appendix 1. Here again, the solution lies in clarifying these differences and the implemented harmonization rules.

Multimedia Appendix 1 gathers complementary discussion points that can help inform interpretation of the harmonized data. Eventually, for complementary knowledge on SIVEP-Malária data quality, readers are encouraged to refer to existing publications on the subject [12,27].

Method Reproducibility

The entire development of the harmonization and visualization applications was carried out with the constant concern that they can be easily and rapidly implemented in other cross-border contexts.

This was ensured by satisfying standards and using existing dedicated and open source tools for data harmonization and visualization. Moreover, the objects of study (ie, patient, consultation, locality, etc) and their properties were formalized by an application knowledge model that currently takes two forms: a dump of the database structure in Structured Query Language (SQL) for its implementation within a database management system such as PostgreSQL, and an ontological formalization in Web Ontology Language (OWL) [28] that enables the knowledge model to be represented according to web data standards and thus ensures its dissemination and reuse by other projects and platforms. Future work will focus on updating and enriching this ontology.

The French Guiana–Brazil cross-border area proved to be an excellent laboratory for the cross-border malaria surveillance issue. It gathers all the specific characteristics of cross-border territories, which make the cross-border malaria issue a major obstacle for the elimination of the disease [2]. The characteristics are as follows: a high diversity of cultures, activities, lifestyles, and languages among the populations; different conceptions, strategies, and means of surveillance, prevention, and control of the disease from one country to another; difficulties in following up with some populations due to their high mobility and possible situations of illegality (ie, undocumented people, illegal activities, etc); and marginalization of border areas with respect to national territorial management and implementation of national public health policies. Moreover, the existing national surveillance systems present significant systemic, syntactic, and semantic differences, and both countries impose different and constraining legal requirements. All the previously listed features make the study area representative of situations we are likely to encounter elsewhere, especially at the international borders of the Brazilian Amazon.

All of the above ensures reproducibility of the method. In fact, the approach was successfully tested at the border between Colombia and Brazil, where a similar monitoring system is currently being developed.

Cross-Border Malaria Dynamics

Interannual dynamics of malaria case numbers result from a conjunction of multiple factors, and it is difficult to state which one is predominant. However, a few suggestions can be made. Thus, the use of RDTs and the introduction of artemisinin-based combination therapies from 2007 in the CDPSs of French Guiana can explain the drop in cases in French Guiana from 2008 [29]. Moreover, in 2008 with the start of the military operation Harpie, which followed operations Anaconda and Toucan, the French army significantly increased pressure on illegal gold mining in French Guiana, expelling more illegal workers, mainly to Brazil, and tending to make illegal gold mining unprofitable. Although there is a delay of one year, this may partly explain the drop in the number of cases reported in French Guiana from 2009 onward, since the gold-miner population represents one of the major *Plasmodium* species reservoirs in French Guiana [30,31].

In 2012, a binational campaign of distribution of long-lasting insecticide-treated mosquito nets (55 mg/m^2 concentration of deltamethrin) was carried out on both sides of the French Guiana–Brazil border, co-conducted by the regional health agency of French Guiana (Agence Régionale de Santé de la Guyane) and the health secretariat of the municipality of Oiapoque in Brazil. This may have contributed to the drop in *P falciparum* cases from 2013.

The recrudescence of the case numbers in 2017 and 2018 is more difficult to explain. In fact, such a recrudescence concerned five countries of the Americas according to the Pan American Health Organization [32]: Brazil, Ecuador, Mexico, Nicaragua, and Venezuela. Brazil reported 174,522 cases between January and November 2017 (ie, 56,690 cases more than for the same period in 2016, which represents a 48% increase) [32]. The Amapá state, meanwhile, has seen the number of cases increase by 23%. French Guiana experienced a significant increase of malaria case numbers for the same period, especially in the municipalities at the border with Brazil [33].

The low number of cases in 2019 can be partly explained by concomitant action-research projects, even if their impacts have still to be evaluated. In 2017 and 2018, the ELIMALAR-PALUSTOP (Elimination of Malaria – Stop Paludisme) project performed an active *Plasmodium* species mass screening by molecular biology—polymerase chain reaction method—among 1566 inhabitants of the Saint-Georges municipality, followed by the treatment of all symptomatic and asymptomatic cases. This should have contributed to the decrease of transmission in this cross-border area. In addition, in 2018 and 2019, the French-Brazilian Malakit project distributed self-diagnosis and self-treatment kits to the gold miners in this cross-border area [34].

Differences in follow-up protocols between French Guiana and Brazil can explain the relatively high number of cases associated with follow-up, possible treatment failures, and relapses in French Guiana. The Brazilian health system involves community health workers who visit patients and help with compliance with treatment. On the other hand, in French Guiana, the health system does not benefit from the action of community health workers. Moreover, Brazil systematically gives primaquine to patients with *P vivax*—except for specific cases including pregnancy—which significantly reduces the risk of relapses, whereas prior glucose-6-phosphate dehydrogenase testing is required in French Guiana, which tends to restrict and delay the use of primaquine [33,35]. This situation makes French Guiana more likely to observe *P vivax* relapses than Brazil. In Brazil, patients with good compliance do not experience relapses; in addition, their follow-up does not require consultations at the health centers and does not generate new notifications in the

Brazilian system. Eventually, such differences can be explained by the fact that the rule for the non-*P falciparum new case* identification implies a longer delay in French Guiana (90 days) than in Brazil (60 days) (see Methods section and Multimedia Appendix 1).

International Cooperation

Partnership was a key factor in the success of the CBMIS development. In fact, an operational multilevel—from local health actors to national organizations—and multidisciplinary partnership, including data science, information systems, epidemiology, parasitology, geography, and geomatics, has been strengthening for about eight years within the framework of several research and regional cooperation programs. Such a partnership is able to mobilize skills and know-how to study other cross-border contexts. The co-construction of the system with all partners ensures its appropriation by health actors so that the system can actually enter into the practice of surveillance and ensure targeted and coordinated public health responses from both countries in order to achieve malaria elimination.

Conclusions

We propose a system that provides comparable and qualified data on the cross-border malaria epidemiological situation. The system is built on technological advances and existing national monitoring systems. Implementing such a system required the application of development good practices, some of which are compulsory, such as those related to privacy, while others contribute to the easy and regular updating of data, facilitate the method's reproducibility, and ensure confidence in the system, thus ensuring the appropriation of results by user communities.

The resulting system is accessible to territory managers, caregivers, researchers, and the general public. The system can notably help in producing new scientific evidence on disease dynamics and determinants, facilitate cross-border cooperation regarding malaria prevention and control, and contribute to citizens' informed participation in public debate and in public authority accountability, in order to achieve malaria elimination.

Acknowledgments

This work was funded by the following entities: the *Fighting malaria: from "global war" to "local guerrillas" at international borders* project, part of the Grand Challenges Explorations Round 18 program funded by the Bill and Melinda Gates Foundation (investment ID OPP1171795); the GAPAM-Sentinela (Guyane Française – Amapá – Amazonas – Malária: Sítio Sentinela Transfronteiriça do Observatório Clima e Saúde) project, part of the Guyamazon program funded by the French National Research Institute for Sustainable Development (Institut de Recherche pour le Développement [IRD]), CIRAD (Centre de Coopération Internationale en Recherche Agronomique pour le Développement), the French Guiana territorial collectivity, the French Embassy in Brazil, FAPEMA (Fundação de Amparo à Pesquisa do Estado do Maranhão), FAPEAP (Fundação de Amparo à Pesquisa do Estado do Amapá), and FAPEAM (Fundação de Amparo à Pesquisa do Estado do Amazonas); the ODYSSEA (Observatory of the Dynamics of Interactions Between Societies and Environment in the Amazon) project, part of the European Union's Horizon 2020 Research and Innovation Program funded by the European Union (Marie Skłodowska Curie grant agreement No. 691053); the Joint International Laboratory (Laboratoire Mixte International [LMI]) *Cross-border observatories of climate, environment and vector-borne diseases - Sentinel site of the Brazilian Climate and Health Observatory* (LMI Sentinela), under the leadership of the IRD, Fiocruz, and Brasilia University; Fiocruz and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Brazil - Finance Code 001; and Santé Publique France. We are also grateful for support from PrInt (Programa de Internacionalização) Fiocruz-CAPES Program.

The authors are grateful to Frédéric Théveny (IRD, Cayenne IRD Center, French Guiana) for his technical assistance in the operational implementation of the system. The authors are also very grateful to Mathilde Isar, Julie Margaillan, and Pierre Bos, from the legal department of the IRD, for their assistance in obtaining the authorization from the French data protection authority (CNIL).

Authors' Contributions

RS and ER wrote the manuscript, with all coauthors commenting on the drafts of the paper; RS also defined the epidemiological indicators, designed and implemented dashboards, ensured SIVEP-Malária data retrieval, and is contributing to the CBMIS maintenance. EM contributed to the harmonization rule definition, the CBMIS evaluation, and the interpretation of the results. CB contributed to the CBMIS conception, the understanding and use of the SIVEP-Malária, and the French-Brazilian scientific cooperation. AC and BG provided the CDPS surveillance system description and contributed to the CBMIS data retrieval. CC installed and is maintaining the CBMIS on the IRD's servers and participated with the development of the CNIL authorization request. JCD participated in the ETL implementation, knowledge formalization, and the CNIL authorization request development. MDSMG, AMM, and PCP contributed to the understanding and use of the SIVEP-Malária and assisted in the geocoding of Brazilian localities and cross-border cooperation. TM contributed to the ETL method implementation. LM contributed to the interpretation of results. AS and BVG supported the cross-border cooperation and participated in the CBMIS evaluation. ER designed and coordinated the project and contributed to the CBMIS design and implementation, to obtaining the CNIL authorization, to the CBMIS maintenance, and to the French-Brazilian scientific cooperation.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Harmonization rules and algorithm (Tables S1 and S2; Figure S1); online dashboard description (Table S3); and percentage of cases with a specified putative infection location in French Guiana, according to epidemiological bulletins of the interregional epidemiology unit of French Guiana (CIRE [Cellule Inter-Regional d'Epidemiologie; Inter-Regional Epidemiological Center]-Guyane/Santé Publique France) (Table S4).
[PDF File (Adobe PDF File), 227 KB-Multimedia Appendix 1]

References

1. World Health Organization. Global Technical Strategy for Malaria 2016-2030. Geneva, Switzerland: World Health Organization; 2015. URL: http://apps.who.int/iris/bitstream/10665/176712/1/9789241564991_eng.pdf?ua=1&ua=1 [accessed 2020-08-11]
2. Wangdi K, Gatton ML, Kelly GC, Clements ACA. Cross-border malaria: A major obstacle for malaria elimination. Adv Parasitol 2015 Jun;89:79-107. [doi: 10.1016/bs.apar.2015.04.002] [Medline: 26003036]
3. Edwards HM, Canavati SE, Rang C, Ly P, Sovannaroth S, Canier L, et al. Novel cross-border approaches to optimise identification of asymptomatic and artemisinin-resistant Plasmodium infection in mobile populations crossing Cambodian borders. PLoS One 2015;10(9):e0124300 [FREE Full text] [doi: 10.1371/journal.pone.0124300] [Medline: 26352262]
4. Krisher LK, Krisher J, Ambuludi M, Arichabala A, Beltrán-Ayala E, Navarrete P, et al. Successful malaria elimination in the Ecuador-Peru border region: Epidemiology and lessons learned. Malar J 2016 Nov 28;15(1):573 [FREE Full text] [doi: 10.1186/s12936-016-1630-x] [Medline: 27894320]
5. Recht J, Siqueira AM, Monteiro WM, Herrera SM, Herrera S, Lacerda MVG. Malaria in Brazil, Colombia, Peru and Venezuela: Current challenges in malaria control and elimination. Malar J 2017 Jul 04;16(1):273 [FREE Full text] [doi: 10.1186/s12936-017-1925-6] [Medline: 28676055]
6. Feachem RGA, Phillips AA, Hwang J, Cotter C, Wielgosz B, Greenwood BM, et al. Shrinking the malaria map: Progress and prospects. Lancet 2010 Nov 06;376(9752):1566-1578 [FREE Full text] [doi: 10.1016/S0140-6736(10)61270-6] [Medline: 21035842]
7. Moonen B, Cohen JM, Snow RW, Slutsker L, Drakeley C, Smith DL, et al. Operational strategies to achieve and maintain malaria elimination. Lancet 2010 Nov 06;376(9752):1592-1603 [FREE Full text] [doi: 10.1016/S0140-6736(10)61269-X] [Medline: 21035841]
8. da Cruz Franco V, Peiter PC, Carvajal-Cortés JJ, Dos Santos Pereira R, do Socorro Mendonça Gomes M, Suárez-Mutis MC. Complex malaria epidemiology in an international border area between Brazil and French Guiana: Challenges for elimination. Trop Med Health 2019;47:24 [FREE Full text] [doi: 10.1186/s41182-019-0150-0] [Medline: 31007535]
9. Institut National de la Statistique et des Études Économiques (INSEE). URL: https://www.insee.fr [accessed 2020-06-12]
10. Instituto Brasileiro de Geografia e Estatística (IBGE). URL: https://www.ibge.gov.br [accessed 2020-06-12]
11. Brazilian Ministry of Health - Health Surveillance Secretariat. Orientações para o preenchimento do SIVEP-Malária. Biblioteca Virtual em Saúde do Ministério da Saúde. 2014. URL: http://bvsms.saude.gov.br/bvs/folder/orientacoes_preenchimento_sivep_malaria.pdf [accessed 2020-06-12]
12. Wiefels A, Wolfarth-Couto B, Filizola N, Durieux L, Mangeas M. Accuracy of the malaria epidemiological surveillance system data in the state of Amazonas. Acta Amazon 2016 Dec;46(4):383-390 [FREE Full text] [doi: 10.1590/1809-4392201600285]
13. Hanf M, Stéphani A, Basurko C, Nacher M, Carme B. Determination of the Plasmodium vivax relapse pattern in Camopi, French Guiana. Malar J 2009 Dec 04;8:278 [FREE Full text] [doi: 10.1186/1475-2875-8-278] [Medline: 19961585]
14. Shneiderman B. The eyes have it: A task by data type taxonomy for information visualizations. In: Proceedings of the IEEE Symposium on Visual Languages. New York, NY: IEEE; 1996 Presented at: IEEE Symposium on Visual Languages; September 3-6, 1996; Boulder, CO p. 336-343. [doi: 10.1109/VL.1996.545307]
15. Chang W, Cheng J, Allaire J, Xie Y, MacPherson J. Shiny: Web application framework for R. The R Project for Statistical Computing. 2018. URL: https://CRAN.R-project.org/package=shiny [accessed 2020-06-12]
16. Saldanha R, Barcellos C, Roux E. Transborder malaria cases: Notified or with place of residence or infection in the municipalities of Oiapoque (BR), Saint-Georges-de-l'Oyapock, Camopi or Ouanary (FR). Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (ICICT/Fiocruz). URL: https://shiny.icict.fiocruz.br/publicirdmalaria/ [accessed 2020-06-12]
17. Saldanha R, Barcellos C, Roux E. Expert online dashboards for cross-border malaria between French Guiana and Brazil. Brazilian Climate and Health Observatory - French Guiana-Amapá surveillance area. URL: https://irdmalaria.icict.fiocruz.br [accessed 2020-06-12]

18. Barcellos C, Roux E, Ceccato P, Gosselin P, Monteiro AM, de Matos VP, et al. An observatory to gather and disseminate information on the health-related effects of environmental and climate change. Rev Panam Salud Publica 2016 Sep;40(3):167-173. [Medline: 27991974]
19. ICICT, Fiocruz, LMI Sentinela. Amapá-French Guiana. Brazilian Climate and Health Observatory. URL: https://climaesaude.icict.fiocruz.br/amapa-guiana-francesa [accessed 2020-06-12]
20. Dell'Erba M, Fodor O, Ricci F, Werthner H. Harmonise: A solution for data interoperability. In: Proceedings of the Second IFIP Conference on E-Commerce, E-Business, E-Government (I3E 2002). Deventer, Netherlands: Kluwer, BV; 2002 Presented at: The Second IFIP Conference on E-Commerce, E-Business, E-Government (I3E 2002); October 7-9, 2002; Lisbon, Portugal p. 433-445. [doi: 10.1007/978-0-387-35617-4_28]
21. Zinszer K, Shaban-Nejad A, Menon S, Okhmatovskaia A, Carroll L, Painter I, et al. Integrated disease surveillance to reduce data fragmentation: An application to malaria control. In: Proceedings of the 2014 International Society for Disease Surveillance (ISDS) Conference. 2015 Feb 26 Presented at: 2014 International Society for Disease Surveillance (ISDS) Conference; December 9-11, 2014; Philadelphia, PA. [doi: 10.5210/ojphi.v7i1.5849]
22. D'Agostino M, Samuel NO, Sarol MJ, de Cosio FG, Marti M, Luo T, et al. Open data and public health. Rev Panam Salud Publica 2018;42:e66 [FREE Full text] [doi: 10.26633/RPSP.2018.66] [Medline: 31093094]
23. Al Manir MS, Brenas JH, Baker CJ, Shaban-Nejad A. A surveillance infrastructure for malaria analytics: Provisioning data access and preservation of interoperability. JMIR Public Health Surveill 2018 Jun 15;4(2):e10218 [FREE Full text] [doi: 10.2196/10218] [Medline: 29907554]
24. Global Health Observatory (GHO) data. World Health Organization. URL: https://www.who.int/gho/database/en/ [accessed 2020-06-12]
25. de Pina-Costa A, Brasil P, Di Santi SM, Pereira de Araujo M, Suárez-Mutis MC, Faria e Silva Santelli AC, et al. Malaria in Brazil: What happens outside the Amazonian endemic region. Mem Inst Oswaldo Cruz 2014 Aug;109(5):618-633 [FREE Full text] [doi: 10.1590/0074-0276140228] [Medline: 25185003]
26. Braz RM, Barcellos C. Analysis of the process of malaria transmission elimination with a spatial approach to incidence variation in the Brazilian Amazon, 2016. Epidemiol Serv Saude 2018 Sep 03;27(3):e2017253 [FREE Full text] [doi: 10.5123/S1679-49742018000300010] [Medline: 30183869]
27. Moreira Braz R, Tauil PL, Faria E Silva Santelli AC, Fernandes Fontes CJ. Evaluation of the completeness and timeliness of malaria reporting in the Brazilian Amazon, 2003-2012. Epidemiol Serv Saude 2016;25(1):21-32 [FREE Full text] [doi: 10.5123/S1679-49742016000100003] [Medline: 27861675]
28. Mandon T, Desconnets JC, Roux E. Data harmonization ontology for cross-border malaria surveillance. BioPortal. 2018 Oct 12. URL: http://bioportal.bioontology.org/ontologies/IRDG [accessed 2020-06-12]
29. Ginouves M, Veron V, Musset L, Legrand E, Stefani A, Prevot G, et al. Frequency and distribution of mixed Plasmodium falciparum-vivax infections in French Guiana between 2000 and 2008. Malar J 2015 Nov 10;14:446 [FREE Full text] [doi: 10.1186/s12936-015-0971-1] [Medline: 26555553]
30. Douine M, Musset L, Corlin F, Pelleau S, Pasquier J, Mutricy L, et al. Prevalence of Plasmodium spp in illegal gold miners in French Guiana in 2015: A hidden but critical malaria reservoir. Malar J 2016 Jun 09;15:315 [FREE Full text] [doi: 10.1186/s12936-016-1367-6] [Medline: 27277831]
31. Pommier de Santi V, Dia A, Adde A, Hyvert G, Galant J, Mazevet M, et al. Malaria in French Guiana linked to illegal gold mining. Emerg Infect Dis 2016 Feb;22(2):344-346 [FREE Full text] [doi: 10.3201/eid2202.151292] [Medline: 26811971]
32. Pan American Health Organization / World Health Organization. Epidemiological Alert: Increase of Malaria in the Americas. Washington, DC: PAHO/WHO; 2018 Jan 30. URL: https://www.paho.org/hq/index.php?option=com_docman&view=download&category_slug=2018-9581&alias=43434-30-january-2018-malaria-epidemiological-update-434&Itemid=270&lang=en [accessed 2020-06-12]
33. Mosnier E, Dusfour I, Lacour G, Saldanha R, Guidez A, Gomes M, et al. Resurgence risk for malaria, and the characterization of a recent outbreak in an Amazonian border area between French Guiana and Brazil. BMC Infect Dis 2020 May 26;20(1):373 [FREE Full text] [doi: 10.1186/s12879-020-05086-4] [Medline: 32456698]
34. Douine M, Sanna A, Galindo M, Musset L, Pommier de Santi V, Marchesini P, et al. Malakit: An innovative pilot project to self-diagnose and self-treat malaria among illegal gold miners in the Guiana Shield. Malar J 2018 Apr 10;17(1):158 [FREE Full text] [doi: 10.1186/s12936-018-2306-5] [Medline: 29631588]
35. Musset L, Pelleau S, Girod R, Ardillon V, Carvalho L, Dusfour I, et al. Malaria on the Guiana Shield: A review of the situation in French Guiana. Mem Inst Oswaldo Cruz 2014 Aug;109(5):525-533 [FREE Full text] [doi: 10.1590/0074-0276140031] [Medline: 25184998]

Abbreviations

CAPES: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Coordination for the Improvement of Higher Education Personnel)
CBMIS: cross-border malaria information system

CDPS: Centre Délocalisé de Prévention et de Soins (Center for Prevention and Care), or Service des Centres Délocalisés de Prévention et de Soins (Department of the Centers for Prevention and Care)
CIRAD: Centre de Coopération Internationale en Recherche Agronomique pour le Développement (French Agricultural Research Centre for International Development)
CNIL: Commission Nationale de l'Informatique et des Libertés (National Commission for Computing and Liberties)
ELIMALAR-PALUSTOP: Elimination of Malaria – Stop Paludisme
ETL: extract, transform, and load
FAPEAM: Fundação de Amparo à Pesquisa do Estado do Amazonas (Amazonas State Research Support Foundation)
FAPEAP: Fundação de Amparo à Pesquisa do Estado do Amapá (Amapá State Research Support Foundation)
FAPEMA: Fundação de Amparo à Pesquisa do Estado do Maranhão (Maranhão State Research Support Foundation)
Fiocruz: Fundação Oswaldo Cruz (Oswaldo Cruz Foundation)
GAPAM-Sentinela: Guyane Française – Amapá – Amazonas – Malária: Sítio Sentinela Transfronteiriça do Observatório Clima e Saúde (French Guiana – Amapá – Amazonas – Malaria: Cross-Border Sentinel Site of the Brazilian Climate and Health Observatory)
IRD: Institut de Recherche pour le Développement (French National Research Institute for Sustainable Development)
LMI: Laboratoire Mixte International (Joint International Laboratory)
LVC: lâmina de verificação de cura (treatment verification slide)
ODYSSEA: Observatory of the Dynamics of Interactions Between Societies and Environment in the Amazon
OWL: Web Ontology Language
PrInt: Programa de Internacionalização (Internationalization Program)
RDT: rapid diagnostic test
SIVEP-Malária: Sistema de Informações de Vigilância Epidemiológica da Malária (Malaria Epidemiological Surveillance Information System)
SQL: Structured Query Language
WHO: World Health Organization

Edited by Z El-Khatib, T Sanchez; submitted 14.07.19; peer-reviewed by C Luz, A Ramachandran; comments to author 20.12.19; revised version received 05.05.20; accepted 01.06.20; published 01.09.20

Please cite as:
Saldanha R, Mosnier É, Barcellos C, Carbunar A, Charron C, Desconnets JC, Guarmit B, Gomes MDSM, Mandon T, Mendes AM, Peiter PC, Musset L, Sanna A, Van Gastel B, Roux E
Contributing to Elimination of Cross-Border Malaria Through a Standardized Solution for Case Surveillance, Data Sharing, and Data Interpretation: Development of a Cross-Border Monitoring System
JMIR Public Health Surveill 2020;6(3):e15409
URL: http://publichealth.jmir.org/2020/3/e15409/
doi: 10.2196/15409
PMID:

4.5 Da ideia, os desafios metodológicos e tecnológicos à visualização de dados

Com o avanço da epidemia de COVID-19, tornando-se uma pandemia e chegando ao Brasil em Fevereiro de 2020, pesquisadores, tecnologistas e bolsistas do ICICT criaram algumas iniciativas para resposta à pandemia. Dentre elas, um sistema de monitoramento da pandemia, que agrega mais de 10 fontes de dados para uma apresentação visual do avanço da pandemia no Brasil e outros países.

O capítulo de livro a seguir apresenta a história da criação deste sistema, os bastidores de uma equipe de *ciência de dados*, suas relações institucionais e pessoais e principais dificuldades na criação e manutenção de um projeto que realiza um ciclo completo, da coleta de dados à divulgação científica e civil de seus resultados.

Este material foi produzido à convite, como um dos capítulos do livro "Cenários Epidemiológicos e Vigilância em Saúde na Covid-19", organizado pelos pesquisadores Dr. Carlos Machado de Freitas, Dr. Christovam Barcellos e Dr. Daniel Antunes Maciel Villela.

O livro irá integrar uma série de *Instant Books* denominada "Informação para Ação na Covid-19", resultado de uma parceria entre o Observatório Covid-19 Fiocruz com a Editora Fiocruz e apoio Scielo Livros, permitindo a ampla divulgação dos livros e sua constituição como uma memória do enfrentamento de emergências em saúde pública futuras, como afirma a carta-convite apresentada logo após o capítulo.

MonitoraCovid-19: informação e disseminação de indicadores em uma pesquisa multidisciplinar

Autores: Raphael Saldanha, Diego Ricardo Xavier, Mônica Magalhães, Paulo Borges, Christovam Barcellos, Marcel Pedroso.

Introdução

O presente artigo apresenta o processo de criação, implementação e manutenção do projeto MonitoraCovid-19, enfatizando as questões organizacionais, científicas e da sociedade civil que o cerca.

O projeto foi construído no Instituto de Comunicação e Informação Científica e Tecnológica (ICICT), da Fundação Oswaldo Cruz (Fiocruz), que tem grande tradição na criação e manutenção de observatórios e sistemas de informática para doenças e agravos, com recortes para informação científica e tecnológica e também sobre comunicação em saúde. O instituto desenvolve estratégias e executa ações de informação e comunicação no campo da ciência, tecnologia e inovação em saúde, objetivando atender às demandas sociais do Sistema Único de Saúde (SUS) e de outros órgãos governamentais.

A missão da unidade se traduz em ações integradas de pesquisa e ensino; comunicação e informação; e gestão e desenvolvimento institucional, alinhadas a quatro eixos temáticos: Desafios do SUS; Ciência e Tecnologia, Saúde e Sociedade; Inovação na Gestão; e Saúde, Ambiente e Sustentabilidade. O objetivo comum desses eixos é fortalecer o SUS e promover melhores condições de vida e saúde da população. O Instituto atua no campo da Informação e Comunicação em Saúde, uma área interdisciplinar do conhecimento, que por sua ampla abrangência gera, ao mesmo tempo, desafios e oportunidades (ICICT, 2020).

O Laboratório de Informação em Saúde (LIS) do ICICT desenvolve e mantém diversos observatórios e sistemas dedicados a monitorar a saúde da população brasileira e de outros países conveniados, acumulando larga experiência na área. Em colaboração com o Ministério da Saúde e com o Instituto Brasileiro de Geografia e Estatística (IBGE), o LIS é responsável pelo aprimoramento do Sistema de Informação sobre Mortalidade, Sistema de Monitoramento da AIDS (MONITORAIDS), e pelo planejamento e desenvolvimento da Pesquisa Nacional de Saúde do IBGE (PNS). O LIS desenvolve sistemas de indicadores que foram criados e são ofertados publicamente com foco no acesso e uso de serviços de saúde pela população brasileira (PROADESS), população de idosos

(SISAP), qualidade da água e saúde (Água Brasil), monitoramento de clima e saúde (Observatório de Clima e Saúde) e acompanhamento de indicadores de mortalidade infantil (MONITORIMI). Estes sistemas dispõem de informações que servem como subsídios para gestores, tomadores de decisão, e também buscam apresentar as informações em linguagens acessíveis à comunidade acadêmica e ao cidadão usuário do SUS.

O LIS abriga o Núcleo de Geoprocessamento, que atua na produção, adequação e atualização de dados e análises espaciais que relacionam dados socioeconômicos, de saúde e ambiente. Seus pesquisadores, tecnologistas e bolsistas estão rotineiramente envolvidos na construção e/ou manutenção de algum destes sistemas, o que permitiu uma acumulação de conhecimentos e técnicas que cobrem diversos aspectos da implementação de observatórios. Nesse contexto surgiu o projeto MonitoraCovid-19 .

Reorganizando os esforços

Em março de 2020, com o avanço da pandemia de Covid-19 no Brasil, o trabalho presencial na Fiocruz começou a ser flexibilizado. As salas e laboratórios passaram a ficar gradativamente mais vazias, com eventos, reuniões e encontros postergados.

Após algumas semanas de adequação ao novo regime de trabalho, a condução dos projetos e pesquisas em andamento foram reorganizadas. Novas datas e prazos foram acordados com instituições parceiras, que invariavelmente também se encontravam na mesma situação.

Reuniões virtuais periódicas começaram a ser agendadas, o fluxo de e-mails aumentou e grupos em aplicativos de conversa se multiplicaram enquanto novas diretivas e planos de contingência eram publicados e atualizados pela Fiocruz e ICICT.

O Brasil e a pandemia

Ainda que distante, em outros continentes, os casos e óbitos de Covid-19 naturalmente captavam a atenção de alguns profissionais do LIS e discussões sobre a pandemia surgiam nas conversas e e-mails, principalmente sobre estimativas, taxas e outros cálculos importantes para COVID-19.

Em seguida, começamos a nos questionar sobre a situação brasileira. Na época, as questões eram “Quando a pandemia irá chegar aqui? Como o governo está se preparando? O que deve ser feito?”

Após 26 de fevereiro de 2020, com os primeiros casos autóctones de COVID-19 no Brasil, alguns cálculos começaram a ser feitos para a realidade brasileira. Acompanhamos pela imprensa os números que timidamente começaram a ser divulgados. Nos meses de fevereiro e março, coletivas do Ministério da Saúde eram diariamente realizadas e transmitidas por praticamente todas as emissoras de televisão. Conforme os números aumentavam, a discussão entre alguns pesquisadores, tecnologistas e bolsistas ficou mais intensa.

Nesta época, o Ministério da Saúde não disponibilizava um repositório de dados sobre casos e óbitos aberto para o público. Algumas iniciativas pontuais para registrar os números apresentados nas coletivas começaram a surgir, sendo divulgadas na Internet em grupos de conversa de jornalistas e cientistas de dados. Pode-se mencionar o *dataset* publicado por Raphael Fontes na plataforma Kaggle[1] e os dados publicados por Wesley Cota em seu repositório do GitHub[2] como algumas das primeiras fontes de dados estruturados e atualizados regularmente sobre a pandemia de COVID-19 no Brasil.

Com a disponibilidade de dados internacionais, como o repositório *do European Centre for Disease Prevention and Control* (ECDC) e dados nacionais por UF, surgiu no ICICT a primeira iniciativa formal de acompanhar os dados da pandemia de COVID-19.

O projeto toma forma

Em meados de março de 2020 foi criada uma planilha no software Excel. Com dados do ECDC, foram produzidos alguns gráficos de casos e óbitos e cálculos de fatores de crescimento comparativos entre países. O projeto nasceu em uma planilha, feita para atender a nossa curiosidade pessoal sobre a evolução da pandemia nos países e sua chegada no Brasil.

Contudo, a atualização da planilha com dados novos e seu incremento com mais funções se mostrou insustentável. A cada dia entravam novos países na lista, as fontes de dados mudavam seus padrões repentinamente, os dados precisavam ser baixados manualmente, convertidos de formato e inseridos na planilha manualmente, as fórmulas precisavam ser atualizadas para a entrada de novas datas. Em função dessa dinâmica, o editor de planilhas estava em seu limite.

[1] https://www.kaggle.com/unanimad/corona-virus-brazil
[2] https://github.com/wcota/covid19br

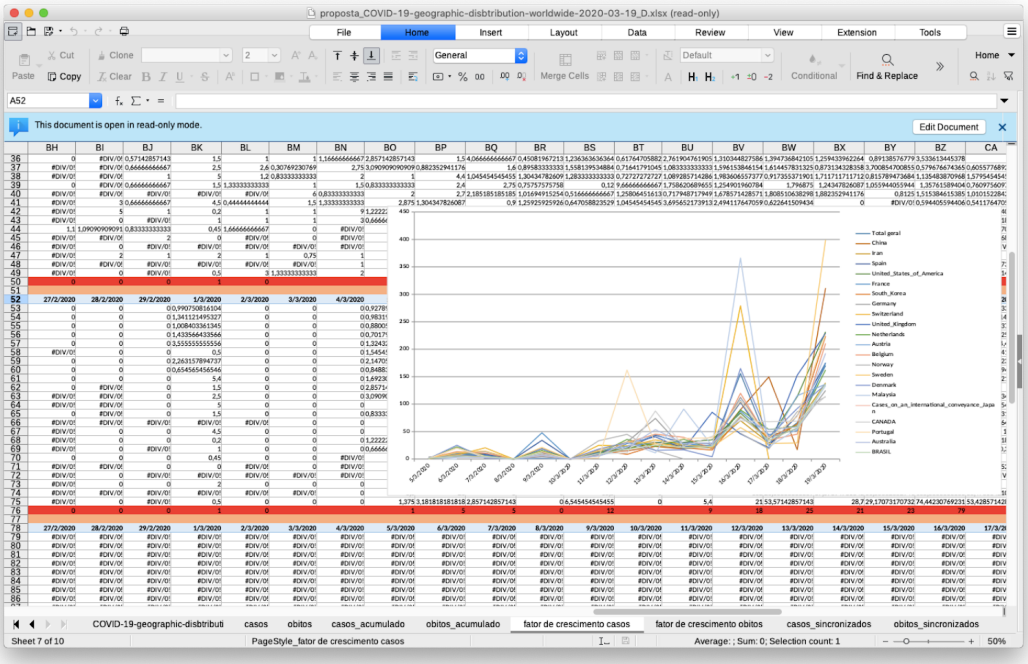


Figura 1: Planilha de cálculos

Neste momento, percebemos que a quantidade de dados aumentava de forma acelerada e que essa seria a tônica da pandemia, o que era só uma curiosidade sobre o processo epidêmico foi se transformando em necessidade de contribuir no enfrentamento do problema de saúde pública. Percebemos nessa época que a pandemia produziria uma grave crise de saúde pública, mas nem nossas piores previsões apontavam para o cenário que se desenharia.

Então, surgiu a ideia de reproduzir todo esse processo manual de maneira mais automatizada. Um programa que conseguisse baixar os dados da Internet, criar uma base única de casos e óbitos por Covid-19 para países e UFs, produzir gráficos e fazer cálculos de forma automática, conforme novos dados fossem disponíveis. Neste momento surge o conceito fundamental do MonitoraCovid-19.

Passamos então a procurar uma ferramenta mais flexível e escalável para a enxurrada de dados e ideias sobre o que fazer com esses dados. O pacote estatístico R nos possibilitou realizar a importação de diversos tipos de dados com mais facilidade e a possibilidade de criar um aplicativo *on-line*.

Uma das primeiras versões do sistema, ainda sem o nome, apresentava apenas 6 abas, conforme a figura 2, de 23 de março de 2020.

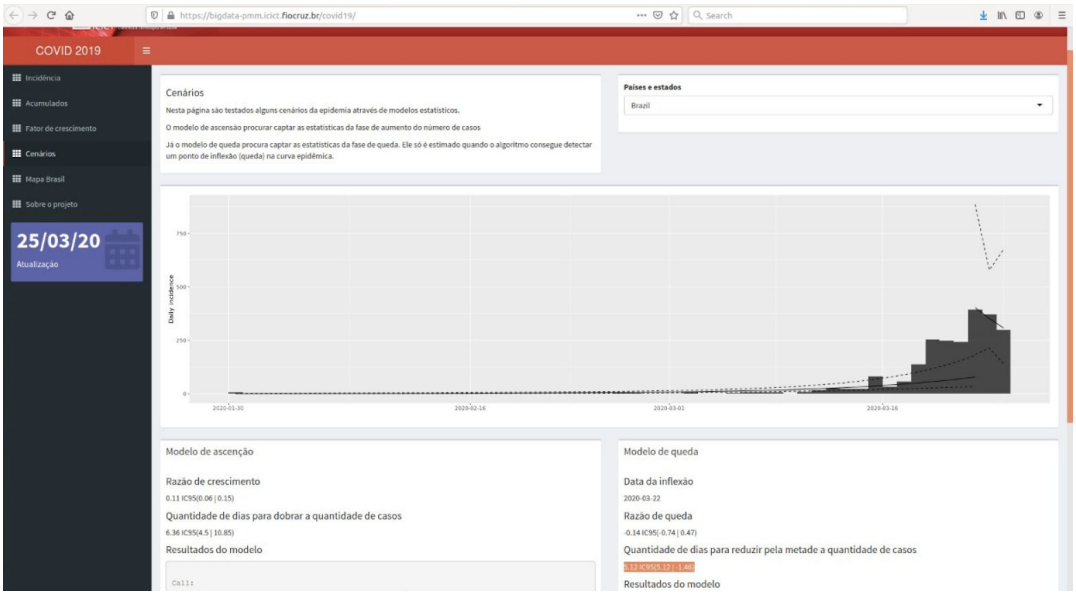


Figura 2: Imagem de uma das primeiras versões do MonitoraCovid-19

De março ao final de novembro de 2020, as atualizações das bases de dados eram feitas manualmente. Durante 253 dias, os dados foram baixados de forma manual de suas fontes, organizados e padronizados diariamente e passaram a alimentar a aplicação web.

Não fazíamos ideia do quão pesada seria essa rotina de atualização. Ela surgiu ingênua, a partir da nossa intenção de ver os dados novos nos gráficos que estávamos criando. Conforme mais pessoas se interessavam em nossos gráficos, ainda dentro LIS, crescia em nós o senso de dever: manter o projeto atualizado e funcionando, todos os dias.

Cumprir este dever impos fardos sobre nós e nossas famílias. Tornou-se rotina familiar ir trabalhar às 20h e “atualizar os dados” à noite, todos os dias, incluindo finais de semana e feriados. Também ao longo do tempo foram surgindo mais ideias para análise e visualização e isso demandou um trabalho adicional de modelagem de dados e programação.

Apenas em dezembro de 2020, uma rotina completamente automática de atualização foi colocada a serviço. Esta etapa será melhor descrita adiante.

Data driven analysis

Ao escrever sobre o projeto, é tentador desejar afirmar que ele foi o fruto esperado de um extenso e detalhado planejamento, onde tudo foi pensado e

antevisto, afirmar que cada aba e conteúdo foi planejado para ser daquele jeito, desde o começo. Mas seria desonesto. O projeto nasceu sem ser um projeto, ele era um teste, um tubo de ensaio, e hoje, mesmo sendo uma plataforma estável e confiável, continua sendo um laboratório para explorar possibilidades de novas análises e dados.

O método cartesiano do planejamento com objetivos, métodos e resultados obviamente tem seu valor. Mas nesse caso, invertemos de alguma maneira essa sequência. Não ter uma ideia formalizada do que queríamos nos possibilitou maior liberdade e criatividade em explorar diversas possibilidades de análise *data driven*[3]. Gastamos obscenas quantidades de tempo em gráficos que não publicamos, reuniões longas sobre temas que não abordamos. Mas arriscamos e não nos arrependemos, pois não foi desperdício, foi aprendizado.

Nome do projeto

À medida que o projeto avançava, ainda que de forma discreta, precisávamos de um nome para registrar um domínio na Internet e podermos nos referir a ele de forma mais precisa. Após algumas sugestões, decidimos por "MonitoraCovid-19".

Na mesma época, surgiu uma outra iniciativa nacional de destaque, também denominada "MonitoraCovid-19". Tratava-se de um aplicativo desenvolvido pelo Comitê Científico de Combate ao Coronavírus do Consórcio Nordeste e capitaneado pelo renomado cientista Dr. Miguel Nicolelis. O uso do mesmo nome pelos dois projetos causou certa confusão na imprensa, dirimida rapidamente.

Fontes de dados

O projeto apoia-se atualmente em mais de 10 fontes de dados diferentes, contemplando desde a quantidade de casos e óbitos de Covid-19, a variáveis ligadas direta ou indiretamente à doença, como por exemplo, mobilidade urbana, população em risco e medidas legislativas.

Atualmente, utilizamos os dados disponibilizados diariamente pela *John Hopkins Coronavirus Resource Center* de casos e óbitos para os países; dados do Ministério da Saúde para casos e óbitos no Brasil, UFs e municípios; projeto InfoGripe para dados de SRAG; dados do Facebook e Universidade de Maryland para pesquisa de sintomas; dados de congestionamentos de trânsito da plataforma *Waze* disponibilizados pelo *IDB Coronavirus Impact Dashboard*;

[3] Foster Provost e Tom Fawcett. Big Data.Mar 2013.51-59. http://doi.org/10.1089/big.2013.1508

dados de mobilidade urbana disponibilizados pelo *Google*; dados de utilização de transporte público disponibilizados pela plataforma *Waze*; e metadados de legislações municipais, estaduais e federais sobre COVID-19 disponibilizados pelo site Leis Municipais.

Algumas intercorrências aconteceram, principalmente devido a dificuldades de acesso a dados de casos e óbitos de COVID-19 para as UFs e municípios brasileiros. No início da pandemia (fevereiro e março de 2020), o Ministério da Saúde informava a população sobre a quantidade de casos e óbitos exclusivamente através de coletivas de imprensa. O Ministério da Saúde disponibilizou os dados de casos e óbitos em pelo menos 7 sistemas diferentes até o momento, apresentados na Figura 3.

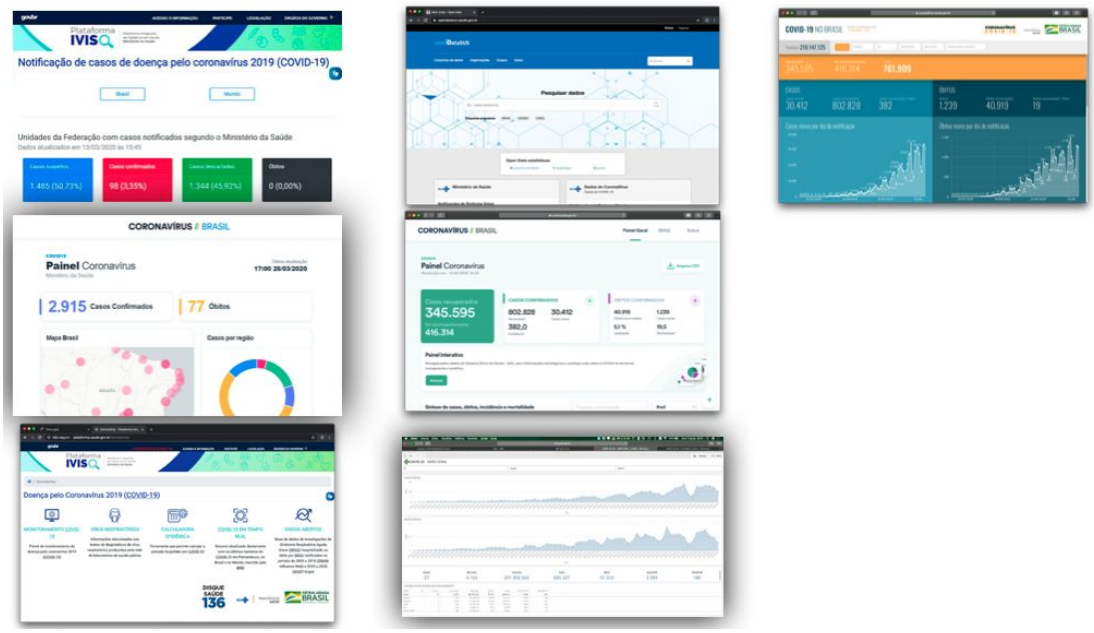


Figura 3: Telas de diferentes sistemas de informação de casos e óbitos de COVID-19 do Ministério da Saúde

Esses sistemas foram apresentados seguidamente pelo Ministério da Saúde, muitas vezes mantendo dois ou três deles em funcionamento e, em algumas vezes, com dados divergentes. Embora alguns sistemas ou versões permitissem o *download* de dados, em vários momentos era oficialmente impossível baixar os dados estruturados oficiais de casos e óbitos por COVID-19 a partir do Ministério da Saúde.

Essas transições entre sistemas e constantes dificuldades em realizar o *download* diários de dados dificultava imensamente o processo de atualização da plataforma. Basicamente, a cada dia podia-se esperar algo diferente, seja na oferta de dados, ou até mesmo variações no formato, extensão, horário de disponibilização dos dados e variáveis constantes nos arquivos. Isto exigia um grande esforço da equipe para manter os dados diários atualizados.

Como não havia um repositório oficial de dados históricos da epidemia no Brasil, dependíamos de projetos e iniciativas que coletavam estes dados diariamente, durante as coletivas, e disponibilizavam os dados históricos estruturados de forma *on-line*.

Em 3 de junho de 2020 ocorreu o chamado "apagão de dados", data em que o painel "Coronavírus Brasil" mantido pelo Ministério da Saúde foi parcialmente descontinuado. Nesta data, regredimos 4 meses em termos de transparência de dados oficiais e passamos a depender diretamente de projetos e iniciativas privadas para manter o projeto com dados atualizados. Passamos então a utilizar diariamente os dados disponibilizados pelo projeto Brasil.IO.

Nos dias seguintes ao apagão, chegamos a até 11.565 usuários em um único dia (7 de junho de 2020), procurando dados atualizados e confiáveis sobre COVID-19 no Brasil.

O projeto Brasil.IO é uma iniciativa de código aberto liderada pelo cidadão Álvaro Justen, reconhecido desenvolvedor nas comunidades de software livre e jornalismo de dados brasileiros. Contando com o trabalho voluntário de dezenas de pessoas, o projeto diariamente acessa os painéis, planilhas, postagens em redes sociais e outras mídias disponibilizadas pelas secretarias estaduais de saúde para reunir os dados de casos e óbitos de Covid-19 no Brasil, para UFs e municípios.

Como contrapartida institucional e reconhecimento do serviço essencial prestado, o MonitoraCovid-19 procurou direcionar parte de seu financiamento e recursos para o projeto Brasil.IO através de doações.

De 3 de junho até 1 de dezembro de 2020, utilizamos diariamente os dados sobre COVID-19 disponibilizados pelo Brasil.IO. A partir de dezembro, notada uma maior estabilidade e qualidade dos dados disponibilizados pelo Ministério da Saúde, retornamos a utilizá-los como fonte primária de dados de casos e óbitos por COVID-19 no Brasil. Essa estabilidade foi interrompida pontualmente no dia 4 de novembro de 2020 por um ataque "hacker" aos servidores do DataSUS. A disponibilização de dados de COVID-19 pelo Ministério da Saúde foi parcialmente interrompida nesta data, sendo plenamente retomada apenas alguns dias depois.

Gerenciando o projeto

Com o início do trabalho remoto, os pesquisadores, tecnologistas e bolsistas do LIS testaram algumas ferramentas de interação: além do e-mail, passamos a usar plataformas de reuniões on-line e, principalmente, grupos de conversa no aplicativo Whatsapp. Além de um grande grupo para interação de todos no laboratório e outro, mais específico para o Núcleo de Geoprocessamento do LIS, criamos alguns grupos para o gerenciamento específico do MonitoraCovid-19.

A participação nestes grupos seguiu uma premissa básica: é incluído no grupo quem se dispor a contribuir, direta e ativamente, na evolução do projeto. Naturalmente, diversas pessoas apoiam o projeto e suas iniciativas, mas a inclusão de todos seria contraprodutiva. Seguimos com a lógica de "*need to know basis*".

Um dos primeiros grupos criados para o projeto denominou-se "Modelagem Covid". Neste grupo, procuramos incluir pessoas do laboratório e colegas de outras unidades e instituições para, inicialmente, conversar sobre modelagem de casos e óbitos de COVID-19, tema a ser detalhado a seguir. Contudo, além deste tema, este grupo também é utilizado para debater diversos assuntos mais amplos do projeto.

Outro grupo criado foi o "Corrigindo atraso". Neste grupo, convidamos colegas da Fiocruz de outro instituto, com expertise nos dados de Síndrome Respiratória Aguda Grave (SRAG) para nos ajudar na construção de uma nota técnica. Atualmente, continuamos debatendo neste grupo sobre os casos de SRAG e outras bases de dados relacionadas.

Por último, foi criado o grupo "Confusão". Como conota o nome, este grupo foi criado para a resolução de assuntos mais urgentes, envolvendo pessoas chave na manutenção do projeto.

Através destes grupos, conseguimos tomar decisões e discutir os assuntos do projeto de forma efetiva, diretamente com os envolvidos no projeto.

Contribuindo no projeto

Essa estrutura colaborativa de gestão e atuação no projeto tem caráter interinstitucional e bastante diferente do organograma da Fiocruz, exigindo convite para colaboração de vários técnicos e pesquisadores, dentro do ICICT e em outros institutos da Fiocruz.

Observamos que, dentro do ICICT, com algumas exceções, poucos pesquisadores puderam colaborar diretamente no desenvolvimento contínuo do projeto. Apesar de enviarem sugestões e pedidos de dados em alguns momentos, outras demandas parecem ter limitado a participação de mais profissionais no projeto. Já fora do ICICT, conseguimos uma colaboração interessante com um pesquisador da Escola Politécnica Joaquim Venâncio, resultando em uma nota técnica publicada e artigos submetidos a periódicos científicos. Em situações pontuais, no entanto, quando solicitados os pesquisadores com expertise em determinados temas, se prontificaram a colaborar. Um exemplo de sucesso foi a parceria entre os pesquisadores para elaboração da nota técnica sobre a volta às aulas, na qual foi estimado a população com fatores de risco e idosa que convive na mesma habitação com pessoas em idade escolar.

Construindo pontes

Ao longo do desenvolvimento do projeto, procuramos agregar novas fontes de dados e análises e, assim, conseguimos estabelecer algumas parcerias, dentro da própria Fiocruz e com outras instituições e empresas.

Na Fiocruz, conseguimos estreitar laços com a equipe do projeto InfoGripe, responsável pela modelagem e monitoramento de dados de SRAG no Brasil. Já conhecíamos parte da equipe em outros projetos e eventos institucionais, mas essa parceria se estabeleceu através de canais bem informais, como mensagens trocadas através do Twitter e WhatsApp. Com a rotina de atualização de dados, passamos a conversar mais intensamente, dirimir dúvidas sobre os dados de SRAG que divulgamos, indicar dados e eventos e propor colaborações entre os projetos.

Nessa ocasião, vários grupos que estudavam a epidemiologia e padrões da Covid-19 se reuniram para discutir os dados e possibilidades de análise. Capitaneado pela equipe do InfoGripe, o webinar “O Panorama da COVID-19 no Rio de Janeiro e Brasil. Onde estamos e para onde vamos?” possibilitou o intercâmbio de informações e alternativas de análise e estudos compartilhados entre diferentes iniciativas.

Figura 4: peça de divulgação do evento “O Panorama da COVID-19 no Rio de Janeiro e Brasil. Onde estamos e para onde vamos?”

O MonitoraCovid-19 também foi apresentado em alguns outros eventos. Pode-se destacar o de 10 de junho, realizado pelo Programa de Pós-Graduação em Informação e Comunicação em Saúde/Icict, intitulado "Acesso aberto a dados de saúde na perspectiva da pandemia"; 19 de junho de 2020 organizado pelo Centro de Estudos do ICICT, intitulado "Informação em Saúde: importância e desafios no enfrentamento da pandemia" e em 4 de agosto de 2020, organizado pelo Fórum de Reportagem sobre a Crise Global em Saúde, intitulado "Como usar o MonitoraCovid-19, sistema da Fiocruz que agrupa dados da pandemia".

Externamente, conseguimos parceria direta com empresas para o fornecimento de dados, como a Moovit, que fornece dados sobre circulação de transporte coletivo nas capitais e regiões metropolitanas. A aplicação avalia globalmente tendências globais de transporte público e combina pesquisas de opinião com dados remotos para construir um retrato de como as pessoas transitam por suas cidades. No período epidêmico da Covid-19, a empresa criou o indicador de porcentagem de redução no uso de transporte público em comparação à média anterior ao período da pandemia.

Usamos, também, dados de acesso público do Google que apontam indicadores de mobilidade da população com dados agregados e anônimos usados em produtos, como o Google Maps. Os indicadores apontam tendências de deslocamento ao longo do tempo por região e em diferentes categorias de locais, como varejo e lazer, mercados e farmácias, parques, estações de transporte público, locais de trabalho e áreas residenciais.

Também incluímos dados do Waze com informações sobre congestionamento de trânsito que podem ser utilizadas para acompanhamento da mobilidade da população[4]. Os dados de variação percentual de quilômetros compartilhados são agregados, anonimizados e foram gerados com base em critérios de privacidade dos usuários. Esses relatórios mostram o aumento ou a redução dos quilômetros rodados como uma variação percentual comparada a dados da linha de base, calculada pela média do dia da semana correspondente ao período de duas semanas antes da decretação de emergência (11 a 25 de fevereiro de 2020).

Também incluímos no sistema dados do Facebook, que disponibiliza indicadores de Covid-19 derivados de pesquisas globais de sintomas em sua plataforma. Os dados são mantidos pela Universidade de Maryland, que os compartilha com outros pesquisadores de saúde. Embora seja uma pesquisa que considera informações individuais, estes dados são disponibilizados de forma agregada o que garante a privacidade do participante. Com base nessas informações, é estimada a porcentagem de pessoas em uma determinada região geográfica em um determinado tempo que relataram sintomas de Covid-19.

Ganhamos apoios de empresas como a RStudio e a Digital Ocean, com cupons de gratuidade para utilização de seus serviços por um tempo limitado. Esta oportunidade foi muito importante para manter em funcionamento o projeto, enquanto seguimos amadurecendo a nossa infra-estrutura de hospedagem do site, tópico que será abordado a seguir.

[4] https://www.waze.com/pt-BR/covid19

Reconhecimento institucional

A resposta da Fiocruz à epidemia se iniciou de forma orgânica, originada em grupos pré-existentes ou que se formaram espontaneamente congregando profissionais e estudantes. À medida em que aumentava a quantidade de casos e óbitos, nos meses de fevereiro e março de 2020, algumas iniciativas surgiram nos diferentes institutos e laboratórios da Fundação.

Visando organizar e institucionalizar essas iniciativas, foi criado na Fiocruz o "Observatório Covid-19: informação para ação"[5]. Após o reconhecimento de cada iniciativa por seu laboratório e instituto, os projetos foram cadastrados junto ao Observatório Covid-19. Em sua página, hospedada no portal Fiocruz, a produção de cada iniciativa é divulgada através de notícias, press releases, notas técnicas, links para seminários e outros materiais. Esta foi uma forma acertada de apoiar e organizar as iniciativas existentes na instituição, evitando hierarquias desnecessárias e sobreposições de trabalhos.

No âmbito do MonitoraCovid-19, a iniciativa foi rapidamente apoiada pelo Núcleo de Geoprocessamento e pela Plataforma de Ciência de Dados Aplicada à Saúde, ambos vinculados ao Laboratório de Informação em Saúde do ICICT. Em seguida, o projeto foi reconhecido como um esforço oficial de resposta à epidemia pelo ICICT, seguindo os trâmites administrativos.

A estratégia de desenvolvimento colaborativo do MonitoraCovid-19, estando sempre aberto à participação de novas pessoas, de dentro e fora da Fiocruz, incluindo voluntários da sociedade civil, levou a discussões sobre seus objetivos e produtos, discussões estas sempre necessárias ao fazer científico. A interação do projeto com outros grupos de pesquisa dentro da Fiocruz foi agitada por sugestões vagas, ou mesmo perturbada por declarações de inutilidade do projeto no início do seu desenvolvimento. Com gradativo amadurecimento da iniciativa, a plataforma MonitoraCovid-19 passou a ser

[5] https://portal.fiocruz.br/observatorio-covid-19

reconhecida com orgulho pela instituição e usada nas apresentações oficiais da Presidência da Fiocruz.

Financiamento do projeto

Para a continuidade sustentável do projeto, procuramos algumas possibilidades de editais de financiamento. Além de recursos financeiros pontuais obtidos com a Presidência da Fiocruz, o projeto foi contemplado pelo edital INOVA Fiocruz "Ideias e Produtos Inovadores - Covid-19 - Encomendas Estratégicas", de 2020. Os editais INOVA da Fiocruz visam financiar projetos internos da instituição, cujo objetivo geral é incentivar ambientes favoráveis à Pesquisa, Desenvolvimento Tecnológico e Inovação em Saúde em todas as áreas de atuação da Fundação Oswaldo Cruz.

Concorrendo com mais de 100 projetos, o MonitoraCovid-19 foi contemplado, recebendo integralmente o valor planejado na proposta. Este recurso tem sido destinado principalmente para o pagamento de bolsas e compra de equipamentos para a manutenção do projeto.

Além do edital INOVA, o projeto foi agraciado com uma premiação da Escola Nacional de Administração Pública (ENAP), na categoria de "Desafios Covid-19", em reconhecimento ao mérito e qualidade do projeto. O recurso tem sido destinado para o pagamento de bolsas e também com auxílio financeiros de projetos de parceiros essenciais para a continuidade do MonitoraCovid-19, bem como de melhoria de infra estrutura na unidade. Cabe registrar que, para o recebimento do prêmio, foi necessário destinar parte de seu valor para arcar com custos administrativos da Fiotec.

A relação com a imprensa

A publicação de notas técnicas e postagens em redes sociais sobre novos gráficos e análises logo atraiu atenção de algumas pessoas e pedidos de entrevistas e solicitação de esclarecimentos sobre a doença e evolução da pandemia começaram a surgir. As demandas da imprensa eram direcionadas

à Assessoria de Comunicação (Ascom) do ICICT, que nos auxiliava na triagem e atendimento desses pedidos. Em outros momentos, éramos diretamente acionados por órgãos de imprensa para comentar um determinado tema em pauta ou produzir indicadores sobre um problema de saúde, sempre usando a plataforma MonitorCovid-19.

Sempre tivemos o entendimento de que o papel prioritário do MonitorCovid-19 era manter a população brasileira bem informada sobre os dados de COVID-19 e o relacionamento com a imprensa era algo essencial.

Em alguns momentos tivemos dificuldade em atender todos os pedidos da imprensa. Além deles, recebemos demandas da imprensa por dados específicos para certas regiões ou municípios. O atendimento a este tipo de demanda era complexo e dividia nossa atenção em manter o projeto sempre atualizado e atender às solicitações. Procuramos seguir considerando que o conhecimento gerado pelo projeto deveria ser divulgado e propor pautas à imprensa.

Até 6 de janeiro de 2021, o portal de notícias do Google apresenta 215 reportagens onde o projeto é citado, em diversos veículos de imprensa, nacionais e internacionais.

E-mails dos usuários

Já no início do projeto, quando a primeira versão foi disponibilizada ao público, uma conta de e-mail institucional foi criada para receber dúvidas, sugestões, críticas e pedidos de esclarecimentos dos usuários.

Este canal de comunicação direta com o projeto se mostrou bastante interessante, ainda que exaustivo em alguns momentos. Recebemos demandas de diversos tipos e origens neste canal, de cidadãos comunicando imprecisões e pedindo esclarecimentos sobre seus municípios, até requisições extraoficiais de dados por entes públicos.

Procuramos responder as mensagens enviadas com rapidez e precisão, mas sem prejudicar o andamento do projeto como um todo. Em alguns momentos, essas demandas por dados e análises específicas foram enviadas pelo Ministério Público.

Acessos ao projeto

O acesso ao projeto foi disponibilizado na Internet no dia 23 de março de 2020. Desta data até o dia 15 de dezembro de 2020, o site do MonitoraCovid-19 teve 211.778 usuários, que abriram as abas do projeto 501.727 vezes (sessões). Atendemos usuários em mais de 100 países de todos os continentes, conforme os dados informados pela plataforma Google Analytics (figura x).

Figura 5a: Dados de acesso ao site do projeto

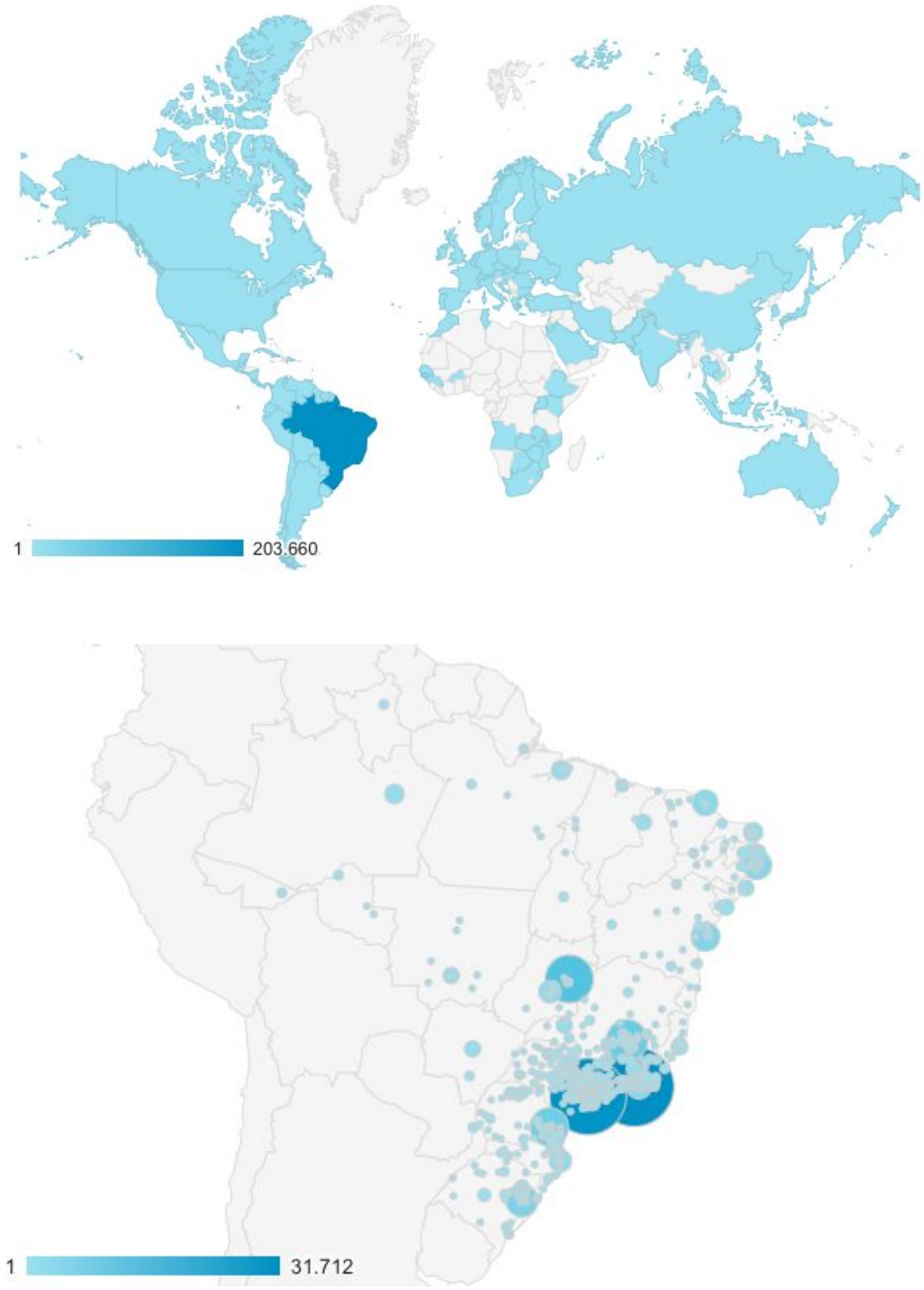


Figura 5b: Dados de acesso ao site do projeto

No Brasil, atendemos usuários em todas as 27 UFs. Pode-se creditar parte destes usuários a divulgação do projeto pelos perfis da Fiocruz nas redes sociais.

Twitter

Conforme o MonitoraCovid-19 avançava com novas funções, bases de dados e gráficos, procurávamos uma forma de divulgar estas novidades. Um dos principais canais que utilizamos é o Twitter. Apesar de não termos uma conta exclusiva para o projeto, procuramos divulgar as novidades do projeto em nossas contas pessoais, já conectadas com diversos usuários e grupos. Em geral, estas postagem alcançam certa repercussão, com curtidas e reportagens por centenas de usuários desta rede social. O Twitter possibilitou também a formação de redes de análise e a troca de informações entre equipes de especialistas de diversas áreas de conhecimento.

Efeito Átila

Um efeito interessante ocorreu quando o pesquisador e divulgador científico Átila Lamarino fez uma postagem em 3 de julho de 2020 sobre o crescimento de casos nas UFs, usando gráficos produzidos no MonitoraCovid-19, incluindo um link para o site do projeto. Esta postagem gerou uma grande quantidade de usuários acessando o site do projeto em um curto espaço de tempo, levando à instabilidade do sistema. Conforme a postagem repercutiu e alcançava mais pessoas, acumulamos um número crescente de usuários. No dia seguinte, atingimos um pico de 3.487 usuários em um único dia.

Infraestrutura de computação e hospedagem

O crescimento do projeto em termos da quantidade de dados trabalhados, gráficos disponíveis e o número de usuários nos deu uma melhor dimensão da responsabilidade ativa do projeto em informar a população, diariamente, sem atrasos ou interrupções.

Para isso, foi necessário rever a infraestrutura utilizada para a hospedagem do projeto. Desde o início, a Plataforma de Ciência de Dados aplicada à Saúde (PCDaS) tem apoiado o projeto MonitoraCovid-19 com sua infraestrutura computacional e revisão técnica do sistema.

Inicialmente, o projeto era hospedado em um servidor da PCDaS localizado fisicamente no Laboratório Nacional de Computação Científica (LNCC), em Petrópolis, RJ. Apesar da alta qualidade ofertada pelo LNCC para computação científica, percebemos que a rede disponível para hospedagem de um site com grande número de acessos e necessidade de alta disponibilidade.

Com o crescimento do número de usuários, optamos então por hospedar o site do projeto dentro da Sala Cofre da Fiocruz, contando com a ajuda da equipe da PCDaS e CTIC/ICICT para a migração e manutenção do sistema. Esta Sala Cofre é um infraestrutura computacional criada na Fiocruz para prover à instituição um ambiente computacional seguro, fisicamente e virtualmente, oferecendo serviços computacionais como hospedagem de sites, com alta confiabilidade e disponibilidade.

Produtos do projeto

Além das abas e gráficos apresentados disponíveis no site, o projeto produziu uma série de notas técnicas, detalhadas abaixo.

A dinâmica da doença demandou maior celeridade na divulgação das informações para que estas pudessem se tornar política pública. Se antes o caminho habitual passava pela construção de um artigo científico e sua publicação para que houvesse alguma visibilidade, durante a pandemia de Covid-19, esse caminho moroso não traria muito ganho do ponto de vista da capacidade de subsidiar decisões e intervenções. Mesmo a velocidade dos *preprints* não possibilitava ganho de oportunidade. A decisão foi: precisamos analisar os dados do sistema e emitir notas técnicas sobre os cenários da pandemia de forma rápida e independente.

Foi o que fizemos. A primeira nota técnica divulgada foi em 2 de abril de 2020 e alertava sobre o processo de interiorização da doença. Apontamos que a doença havia chegado pelas grandes cidades e não existia razão nenhuma para acreditar que não avançasse para as cidades do interior. Foi o que aconteceu. A parceria com o IBGE possibilitou acesso aos dados da pesquisa Regiões de Influência das Cidades (REGIC, edição 2018), em especial, do componente saúde, que foi liberado para uso antes mesmo da sua publicação em julho de 2020. Com base nesses dados, estruturamos outras notas e apontamos o caminho esperado da difusão espacial do vírus.

Em maio de 2020, buscamos avaliar a velocidade de disseminação da doença, como já imaginávamos a chegada da doença no interior do país e historicamente a pouca disponibilidade de recursos para atendimento de alta complexidade (UTI) traria um quadro de desassistência à saúde. Levantamos esses dados e estimamos a distância de deslocamento em busca de atendimento e a disponibilidade de recursos dentro das Regiões de Saúde e o fluxo de pacientes que já estava ocorrendo, de um município em busca de outro município que pudesse atender. Os dados revelaram uma situação trágica que infelizmente se confirmou. Ainda em maio, outra nota avaliava os dados de SRAG do SIVEP-Gripe e como esse sistema permitiria avaliar aspectos como sexo, raça, idade, atendimento, entre outros, e criava um espectro mais amplo para a avaliação da doença.

Em junho de 2020 começou a ocorrer um movimento, com origem em setores da economia e política, para a flexibilização das medidas de isolamento e distanciamento social. Esse movimento, junto a outros processos de desgaste de estratégias de proteção da saúde, resultou na demissão do ministro Mandetta em abril e do ministro Teich em maio. Nesse período nos dedicamos a trabalhar com dados de mobilidade urbana apontando os índices de mobilidade disponíveis no sistema. Também buscamos explicar que existiam tempos epidêmicos diferentes em função da chegada da doença. À medida que a doença chegava, os chamados “epicentros” da doença iam se alterando e

uma flexibilização geral defendida pelo governo federal era uma decisão equivocada.

No mês de julho de 2020, passamos a avaliar os impactos indiretos, sobretudo os óbitos por outras causas e indiretamente ocorridos pela COVID-19. Outros sistemas estavam disponibilizando os dados de cartórios do registro civil, até implementamos essas informações no sistema, mas por conta de inconsistências que observamos à época, bem como distorções e fake news criadas com base nessas informações, retiramos esses dados do sistema. Avaliamos então os padrões de mortalidade no município do Rio de Janeiro, que possui seu próprio site de disponibilização de dados, mostrando um grande volume de óbitos em domicílios, vias públicas e outras unidades de saúde e sem assistência médica, o que poderia estar se repetindo em outras cidades.

Ainda em julho de 2020, a questão da volta às aulas ganhou força no processo de relaxamento das medidas de isolamento social. Nesse período era preocupante essa medida frente à alta da taxa de contágio e, principalmente por conta da exposição da população idosa e da população com fatores de risco que conviviam com pessoas em idade escolar. Fizemos um levantamento com a equipe que detém conhecimento dos dados da Pesquisa Nacional de Saúde e apontamos que quase 10 milhões de pessoas idosas ou com fatores de risco conviviam com crianças em idade escolar. Buscamos ser taxativos, apontando que a retomada das aulas sem testagem, rastreio e às medidas de distanciamento físico, de higiene e o uso de máscara seria o fim do isolamento social para essa população.

Os boletins epidemiológicos das Secretarias Estaduais de Saúde (SES) apresentaram, em alguns momentos, represamento de casos para além do esperado no comportamento cíclico das informações em função dos fins de semana. Decidimos buscar os sistemas de referência (e-SUS-VE e SIVEP-Gripe) para comparar a data do evento (óbitos ou data dos primeiros sintomas nos casos) e a data de divulgação dos dados pelas SES. A diferença entre os

cenários que a população acompanhava na mídia e aqueles baseados na data em que de fato ocorriam os eventos, para alguns estados chegava a quase 2 meses. Essa nota foi lançada em agosto de 2020. Em setembro, em um evento com outros grupos de análise sobre Covid-19 do Brasil, elaboramos uma carta aberta que buscava alertar a mídia e a sociedade sobre essa situação. Concordamos que não era um processo simples e como já havia o hábito da divulgação dos dados pelos boletins, isso não deveria ser alterado, mas complementado. A divulgação continuou sendo pautada pela mídia com os dados dos boletins das SES e continuamos sujeitos a esse represamento, que foi acentuado próximo ao período eleitoral e nos recessos de fim de ano.

Em dezembro de 2020, retomamos a análise sobre excesso de óbitos no município do Rio de Janeiro e os números revelaram um cenário trágico de desassistência na cidade, que com certeza se repetiu em muitas outras cidades. O aumento de casos no fim do ano e maior mobilidade das pessoas motivou a nota técnica que apontava o fim do processo de interiorização e a sincronização da pandemia no espaço. O estudo mostra duas fases na dinâmica espaço-temporal da Covid-19: se em abril e maio tínhamos um processo de difusão da doença, das capitais para cidades de menor porte, e aumento de casos à medida em que a doença avançava no território, já em novembro, a doença ocupava todo o território e a movimentação das pessoas entre cidades definiria o aumento de casos de forma sincronizada em vários locais ao mesmo tempo.

Modelos de previsão

Conforme os casos avançavam no Brasil, em meados de março de 2020, e os primeiros óbitos começavam a ser reportados, diversas propostas nacionais e internacionais surgiram para tentar prever quando ocorreria o auge da quantidade de casos, o chamado "pico da curva", e qual seria a quantidade de casos nesta data. Essas informações seriam de grande importância para

preparar o sistema de saúde, com a preocupação dominante à época de “achatar a curva”, para melhor atender a população.

Estas propostas em geral utilizaram métodos de modelagem epidemiológico compartimentais, como os modelos SIR (Suscetível - Infectado - Recuperado) e algumas variantes.

No âmbito do MonitoraCovid-19, inicialmente procuramos desenvolver modelos de previsão e chegamos a discutir alguns modelos de previsão a curto prazo considerando o crescimento exponencial da doença. Contudo, no decorrer do processo epidêmico, observamos que a epidemia era muito mais complexa do que se esperava e vários dos modelos propostos inicialmente, começaram a ser descartados, pois a incerteza nos modelos para efeitos de médio e longo prazo traziam uma enorme incerteza, devido à falta de conhecimento sobre a doença.

Além disso, enfrentamos - como todos os outros proponentes de modelos - grandes dificuldades com a qualidade e limitações dos dados. Os dados disponíveis sobre a epidemia versavam basicamente sobre a quantidade de casos e óbitos por data de notificação, o que acaba por violar alguns dos pressupostos do modelo. Seria necessário obter a data dos primeiros sintomas dos casos, conhecer precisamente a quantidade de testes aplicados nas populações, detalhes sobre as medidas de prevenção e combate à epidemia e outros fatores que não estavam disponíveis como dados (e não estão até o momento). Por estes motivos, decidimos não apresentar um modelo publicamente. Priorizamos então por trabalhar em análises de curto prazo, sobretudo sobre o processo de disseminação espacial da doença e das implicações relacionadas ao atendimento.

De fato, se observou que diversos dos modelos apresentados publicamente foram discretamente retirados, conforme a epidemia avançava. Em geral, eles eram reajustados diariamente com os novos dados, o que acabava por sempre postergar o "pico de casos". Os valores resultantes desses modelos não

pareciam ser razoáveis com a realidade do momento, ou necessitavam de dezenas de parâmetros adicionais, fornecidos pelo usuário para modelar a curva epidêmica, o que virtualmente impedia qualquer previsão. De fato, nenhum modelo de previsão alertou para a possibilidade futura de um segundo pico de casos, realidade que começou a se definir em novembro de 2020.

Fornecimento de dados para outros projetos

O projeto também fornece os dados para outros projetos e instituições. O IBGE tem um link direto de acesso aos nossos dados atualizados para alimentar seu portal sobre a Covid-19[6]. Outra iniciativa de cessão de dados é para o monitoramento da epidemia na Região Metropolitana de São Paulo.

A estrutura de código

A estrutura de código de programação do projeto se divide em duas partes, *Extraction, Transform and Load* (ETL) e o aplicativo web.

ETL

A estrutura de ETL é responsável pela atualização rotineira dos dados do projeto. Atualmente ela é realizada de forma automática em um servidor e seus passos são notificados via mensagens no Telegram para a equipe técnica do projeto. A execução do script de ETL dura atualmente em torno de uma hora para ser executada.

Esta estrutura é composta de uma série de scripts em linguagem R. O primeiro bloco de scripts realiza o download dos dados utilizados no projeto. Outros scripts efetuam transformações e cálculos necessários, enquanto o último bloco de scripts realiza exportações de dados para que projetos parceiros possam consultar nossos dados.

Aplicativo

[6] https://covid19.ibge.gov.br/

O site do projeto é um aplicativo também desenvolvido com a linguagem R, utilizando-se o pacote Shiny para construção de aplicações online. Trata-se de um arquivo único contendo em torno de 4.000 linhas, onde são definidas a estrutura visual do site, assim como a execução de códigos dinâmicos para a construção das tabelas, gráficos e mapas apresentados no site.

O que vem pela frente

O projeto pretende continuar cumprindo a missão institucional do ICICT, disponibilizando informação de forma aberta e criando interfaces para consultas interativas e aquisição dos dados de forma simples e acessível para a população, os gestores e a sociedade civil.

No curto prazo, o sistema passará por uma remodelagem de layout e usabilidade que pretende tornar a plataforma ainda mais simples e intuitiva tornando-a acessível a um público mais amplo.

Nesse sentido, esperamos incluir dados sobre Covid-19 dos demais sistemas de informação de saúde (SIM, SIH, SIA, etc.) que devem disponibilizar esses dados seguindo uma defasagem inerente a cada sistema de informação, mas que trazem outras possibilidades de análise e entendimento da doença e do processo epidêmico.

As bases de dados do SIVEP-Gripe e e-SUS VE serão trabalhadas para criar interfaces de consultas amigáveis e acessíveis a qualquer usuário sem necessidade de conhecimento especializado sobre modelagem de bancos de dados. Além disso, esperamos incluir os dados sobre vacinação disponíveis tanto no Brasil quanto no mundo.

Esperamos expandir as parcerias e fornecer informação para estudos aprofundados sobre a Covid-19 que nos ajudem a entender a epidemia, para enfrentá-la e criar conhecimento de base para eventuais novos processos endêmicos que venham a ocorrer. Nesse sentido, espera-se maior aproximação das instituições de ensino e das secretarias de saúde para

colaborar na melhoria da informação e na capacitação de pessoal para análise.

REFERÊNCIAS BIBLIOGRÁFICAS

ICICT. Instituto de Comunicação e Informação Científica e Tecnológica em Saúde. Disponível em: https://www.icict.fiocruz.br/content/sobre-o-icict. Acessado em: 20/12/2020.

CENÁRIOS EPIDEMIOLÓGICOS E VIGILÂNCIA EM SAÚDE NA COVID-19

Convite para capítulo do livro instantâneo e do momento

Rio de Janeiro, 18 de novembro de 2020

Prezado **Raphael Saldanha**,

Estamos lhe convidando para que contribua com um capítulo para o "Instant Book" Cenários Epidemiológicos e Vigilância em Saúde na Covid-19.

A proposta é que, junto com coautores escolhidos, você contribua com um capítulo tendo como tema "**A criação e desenvolvimento do painel MonitoraCovid-19**".

O livro instantâneo e do momento Cenários Epidemiológicos e Vigilância em Saúde na Covid-19 integra uma série de Instant Books denominada Informação para Ação na Covid-19, resultado de uma parceria entre o Observatório Covid-19 Fiocruz com a Editora Fiocruz e apoio Scielo Livros, permitindo a ampla divulgação dos livros e sua constituição como uma memória do enfrentamento de emergências em saúde pública futuras.

Todos os capítulos do livro Cenários Epidemiológicos e Vigilância em Saúde na Covid-19 terão como base documentos, notas técnicas, informes, relatórios e seminários já produzidos durante a pandemia, que devem ser estruturados do seguinte modo:

1) Uma introdução situando as principais questões, perguntas e temas que motivaram a elaboração do documento, nota técnica ou relatório quando foi produzido e qual o contexto epidemiológico do mesmo.

2) Conclusões ou considerações finais que apontem para que lições podem ser extraídas das questões, perguntas ou temas contidos no documento, nota técnica, informe ou relatório para o enfrentamento de futuras emergências em saúde pública.

3) A íntegra dos capítulos deverá ter entre 3000 e 4000 palavras, fonte Arial, corpo 12, espaçamento entre linhas 1,5.

4) Devem conter até cinco quadros, gráficos e/ou figuras.

5) O prazo para envio do capítulo é dia 20 de dezembro.

Padrão para referências (exemplos)

Livro

RIVERA, F. J. U. Análise Estratégica em Saúde e Gestão pela Escuta. Rio de Janeiro: Editora Fiocruz, 2003.

Observação: em obras mais de três autores, registrar o nome do primeiro acompanhado do termo et al.

Capítulo de livro

CAPELLÀ, D. & LAPORTE, J. R. Métodos empregados em estudos de utilização de medicamentos. In: LAPORTE, J. R.; TOGNONI, G. & ROZENFELD, S. (Orgs.). Epidemiologia do Medicamento: princípios gerais. São Paulo: Hucitec, Abrasco, 1989.

Artigo de periódico

LESER, W. Crescimento da população e nível de vida em São Paulo. Problemas Brasileiros, 12(134): 16-29, 1974.

Observação: títulos de periódicos não devem ser abreviados. Exemplo: American Journal of Epidemiology.

Monografias, teses e dissertações

TRAIMAN, P. Aspectos Anatômicos da Glândula Lacrimal e de sua Inervação no Macaco-prego Cebus appela, 1988. Dissertação de Mestrado, São Paulo: Instituto de Biociências, Universidade Estadual Paulista.

Figuras, imagens e fotografias devem ser entregues junto com os originais aprovados para publicação (em sua versão final ou definitiva), de acordo com as seguintes especificações:

Figuras, gráficos, tabelas e quadros devem estar preparados em Word/Excel, Ilustrator ou Corel (em vetor ou em .EPS) em arquivos editáveis, sem tabulações, em preto e branco e/ou escala de cinza. No caso de publicação colorida, podem vir em CMYK.

Imagens – fotografias e ilustrações digitalizadas de obras de terceiros – devem ser encaminhadas em formato .TIF ou .JPG, com resolução ideal de 600 DPI, no tamanho a ser reproduzido (mínimo de 10x15 cm), ou maior; quanto às cores, podem estar em escala de cinza (grayscale), CMYK ou RGB.

Todas as figuras devem incluir autoria e fonte.

Atenciosamente,

Carlos Machado de Freitas, Christovam Barcellos & Daniel Antunes Maciel Villela

(organizadores)

5 Conclusão

Religion is a culture of faith; science is a culture of doubt.

—Richard Feynman

As possibilidades de avanços dos métodos científicos, quase sempre, são recebidas com certa incredulidade ou descrédito por alguns pesquisadores. As possibilidades anunciadas por novos métodos costumam ser diminuídas, procura-se intencionalmente igualá-las a métodos já estabelecidos e são rigorosamente postas à prova.

Este processo de crítica e desconfiança de novos métodos, apesar da aspereza e aridez dos debates, é essencial à ciência. Pode-se ponderar – com algumas exceções – que todo novo método nasce puro, com a solene intenção de responder algum tipo de problema ou desafio metodológico, por vezes fruto de alguma pesquisa que sequer visava o desenvolvimento de um método novo método. Contudo, a utilização prática de novos métodos tende a forçar os pré-supostos metodológicos, sendo empregada em campos e aplicações não antes antevistas.

A *ciência de dados* constitui um exemplo de interesse recente. Sua origem não é puramente acadêmica, mas com forte participação do mercado. Negócios, empresas e indústrias criaram a demanda pela análise de grandes massas de dados, não ainda atendidas pelos métodos, *softwares* e *hardwares* disponíveis. Alianças do mercado com a academia possibilitaram o avanço da ciência para atender estas demandas.

Sendo rapidamente percebida por *think tanks* e revistas ligadas à negócios, este movimento é publicizado. O termo *big data* surge e se estabelece como um grande guarda-chuva para todo e qualquer avanço moderno na área de análise de dados.

Gradativamente e com receio, a academia se aproxima ao *big data*, seja pela demanda do mercado por profissionais capacitados, ou pela natural curiosidade científica. O receio é importante: métodos que prometem possibilidades infinitas, grandes fanfarras em publicações e simples modismo já foram vistas antes nas ciências, e poucos destes métodos prevaleceram.

Ao ser recebido pela academia, o termo *big data* é, de certa forma, rebatizado. Passa a se denominar *ciência de dados* e toma vulto de campo um de conhecimento. O *big data* em si, torna-se apenas uma propriedade dos dados, relacionada ao tamanho. O formalismo científico é aplicado, evoluindo a solução comercial de análise de grandes massas de dados para um campo do saber científico.

Na Saúde, a desconfiança e resistência ao *big data* também foi percebida. A análise de dados de saúde é antiga, como foi visto, e a incorporação de novos métodos não poderia

ser fácil. Mas, com o advento da *ciência de dados* na acadêmia, estes novos métodos de grande potencial estão sendo gradativamente admitidos e reconhecidos como válidos, internacionalmente e no Brasil.

Nesta tese, foram apresentados diversos aspectos e potencialidades de *ciência de dados* aplicados à Saúde, possibilitando o estudo do ciclo de geração e disseminação de informação em saúde. Partindo de uma perspectiva histórica, a relação entre dados e saúde foi visitada, apresentando um novo paradigma da *ciência de dados em saúde*, considerando as possibilidades hibridas de uma ciência *theory & data driven* para a Saúde Pública.

Os métodos e modelos de processo de *ciência de dados*, especificamente o KDD, SEMMA e CRISP-DM, foram explorados criticamente. Avaliando seus pontos em comum, um novo modelo de processos denominado KDD-PH foi proposto, sugerindo etapas específicas para um modelo de processos de *ciência de dados* para a pesquisa em saúde pública.

A construção desta tese se deu durante o curso de doutorado em Informação e Comunicação em Saúde, mas se iniciou antes. Conhecimentos prévios – formais e informais – em Geografia, Estatística, Ciências da Computação e Saúde Coletiva foram essenciais para a construção e amadurecimento dos propósitos deste trabalho, de natureza inerentemente interdisciplinar.

O ambiente de livre incentivo à pesquisa, promovido pela Fundação Oswaldo Cruz e, em especial, pelo ICICT, também se mostrou essencial para a produção deste trabalho, oferecendo terreno fértil para introdução e testagem de novos métodos e estratégicas parcerias nacionais e internacionais para o amadurecimento e intercâmbio de saberes.

O formato de construção e apresentação da tese foi apropriado, conjugando capítulos teóricos sobre revisões históricas e métodos, com artigos científicos que se propuseram a colocar em prática tais métodos frente ao escrutino de pares.

O primeiro objetivo específico desta tese versa sobre o estudo da evolução do tratamento e visualização de dados em saúde. O capítulo Referencial teórico (p. 19) oferece uma perspectiva histórica da relação entre dados em saúde, enquanto que a seção Construção teórica (p. 64) apresenta, como um resultado desta tese na forma de artigo, uma análise atual das contribuições de *ciência de dados* para os campos da saúde e demografia. O segundo objetivo específico propõe o acompanhamento e descrição de processos modernos de tratamento e visualização de dados. O capítulo Metodologia (p. 40) se propõe a revisar os principais modelos de processos de *ciência de dados*, propondo por fim um novo modelo de processos, dedicado à pesquisa em saúde pública.

Já o terceiro objetivo específico propõe a aplicação de modelos de processos de análise de dados em pesquisas da área de saúde. Os resultados na forma de artigos apresentados nas seções Captura e seleção de dados (p. 80), Do constructo teórico à

mineração (p. 94), Da coleta distribuída à visualização (p. 109) e Da ideia, os desafios metodológicos e tecnológicos à visualização de dados (p. 125) contemplam diversos passos do modelo de processos KDD.

Por fim, o quarto objetivo, que convida uma análise crítica das vantagens e limites do uso das metodologias de *ciência de dados*, também é abordado no artigo apresentado na seção Construção teórica (p. 64).

As possibilidades de análise de dados não estruturados, como postagens em redes sociais *on-line* e prontuários médicos, tão como certos aspectos éticos e morais de *ciência de dados em saúde* não foram abordadas especificamente nesta tese e devem ser tratados em trabalhos futuros.

A *ciência de dados em saúde* é um campo novo, que oferece novos métodos para responder perguntas que muitas vezes ainda não foram feitas, conjugando novos e antigos métodos. Os interesses em sua aplicação são diversos, da pesquisa acadêmica à análise atuária de mercado. Contudo, sua motivação deve permanecer sólida: promover amplamente a saúde, em dua definição mais vasta, para a população.

Referências

AALST, W. van der. *Process Mining*. 2. ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2016. v. 0. ISBN 978-3-662-49850-7. Citado na página 14.

AGUIAR, F. P. et al. Confiabilidade da informação sobre município de residência no Sistema de Informações Hospitalares - Sistema Único de Saúde para análise do fluxo de pacientes no atendimento do câncer de mama e do colo do útero. *Cadernos Saúde Coletiva*, v. 21, n. 2, p. 197–200, jun 2013. ISSN 1414-462X. Disponível em: <http://www.scielo.br/scielo.php?script=sci{_}arttext{&}pid=S1414-462X2013000200015{&}lng=p>. Citado na página 46.

ALMEIDA-FILHO, N. de. Bases históricas da Epidemiologia. *Cad. Saúde Pública*, v. 2, n. 3, p. 304–311, 1986. Citado na página 14.

ALMEIDA, M. B. Uma abordagem integrada sobre ontologias: Ciência da informação, ciência da computação e filosofia. *Perspectivas em Ciencia da Informacao*, v. 19, n. 3, p. 242–258, 2014. ISSN 19815344. Citado 2 vezes nas páginas 53 e 54.

ALMEIDA, M. F. de. Descentralização de Sistemas de Informação e o uso das informações a nível Municipal. *Informe Epidemiológico do Sus*, v. 7, n. 3, p. 27–33, sep 1998. ISSN 0104-1673. Disponível em: <http://scielo.iec.pa.gov.br/scielo.php?script=sci{_}arttext{&}pid=S0104-16731998000300003{&}lng=en{&}nrm>. Citado na página 46.

ALMEIDA, M. F. de; ALENCAR, G. P. Informações em saúde: Necessidade de introdução de mecanismos de gerenciamento dos sistemas. *Informe Epidemiológico do Sus*, v. 9, n. 4, p. 241–249, dec 2000. ISSN 0104-1673. Disponível em: <http://scielo.iec.pa.gov.br/scielo.php?script=sci{_}arttext{&}pid=S0104-16732000000400003{&}lng=pt{&}nrm>. Citado na página 43.

ANDERSON, C. The end of a theory: the data deluge makes the scientific method obsolete. *Wired*, 2008. Disponível em: <https://www.wired.com/2008/06/pb-theory/>. Citado na página 38.

ARAÚJO, I. S. de; CARDOSO, J. M. Comunicação e saúde. In: FIOCRUZ (Ed.). *Dicionário de Educação Profissional em Saúde*. 2. ed. Rio de Janeiro: Fiocruz, 2007. p. 152. ISBN 978-85-7541-125-4. Disponível em: <http://www.epsjv.fiocruz.br/dicionario/verbetes/comsau.html>. Citado na página 52.

AZEVEDO, A.; SANTOS, M. KDD, SEMMA and CRISP-DM: A parallel overview. *IADIS European Conference Data Mining*, p. 182–185, 2008. Citado 4 vezes nas páginas 54, 56, 57 e 58.

BASKETT, L.; LEROUGE, C.; TREMBLAY, M. C. Using the dashboard technology properly. *Health progress*, v. 89, n. 5, p. 16–23, 2008. ISSN 0882-1577. Citado na página 51.

BITTENCOURT, S. A.; CAMACHO, L. A. B.; LEAL, M. d. C. O Sistema de Informação Hospitalar e sua aplicação na saúde coletiva. *Cadernos de Saúde Pública*, v. 22, n. 1, p. 19–30, jan 2006. ISSN 0102-311X. Disponível em:

<http://www.scielo.br/pdf/csp/v22n1/03.pdfhttp://www.scielo.br/scielo.php?script=sci{_}arttext{&}pid=S0102-311X2006000100003{&}lng=p>. Citado na página 43.

BOERMA, T.; STANSFIELD, S. K. Health statistics now: are we making the right investments? *Lancet*, v. 369, n. 9563, p. 779–86, mar 2007. ISSN 1474-547X. Disponível em: <http://www.sciencedirect.com/science/article/pii/S014067360760364X>. Citado na página 33.

BOILSON, A. et al. Transforming Health through Big Data: Challenges and Considerations. *UK Academy for Information Systems Conference Proceedings*, v. 12, 2018. Disponível em: <https://aisel.aisnet.org/ukais2018/12>. Citado na página 16.

BOYD, D.; CRAWFORD, K. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information Communication and Society*, v. 15, n. 5, p. 662–679, 2012. ISSN 1369118X. Citado 2 vezes nas páginas 37 e 38.

BOYD, D. M.; CRAWFORD, K. Critical Questions for Big Data. *Communication & Society*, v. 15, n. 5, p. 662–679, 2012. ISSN 1369-118X. Citado na página 36.

BRANCO, M. A. F. Sistemas de informação em saúde no nível local. *Cadernos de Saúde Pública*, v. 12, n. 2, p. 267–270, jun 1996. ISSN 0102-311X. Disponível em: <http://www.scielo.br/scielo.php?script=sci{_}arttext{&}pid=S0102-311X1996000200016{&}lng=p>. Citado na página 46.

BRAZ, R. M. et al. Avaliação da completude da variável raça/cor nos sistemas nacionais de informação em saúde para aferição da equidade étnico-racial em indicadores usados pelo Índice de Desempenho do Sistema Único de Saúde. *Saúde em Debate*, v. 37, n. 99, p. 554–562, 2014. Citado na página 46.

CASTELLANOS, P. L. *Sobre o Conceito de Saúde-doença . Descrição e Explicação da Situação de Saúde*. [S.l.], 1990. Citado na página 60.

CHAPMAN, P. et al. *CRISP-DM 1.0 Step-by-step data mining guide*. [S.l.], 2000. Disponível em: <http://www.crisp-dm.org/CRISPWP-0800.pdf>. Citado na página 36.

CHENG, C. K. et al. Digital Dashboard Design Using Multiple Data Streams for Disease Surveillance With Influenza Surveillance as an Example. *Journal of Medical Internet Research*, v. 13, n. 4, p. e85, oct 2011. ISSN 1438-8871. Disponível em: <http://www.jmir.org/2011/4/e85/>. Citado na página 51.

CHIAVEGATTO-FILHO, A. D. P. Uso de big data em saúde no Brasil: perspectivas para um futuro próximo. *Epidemiologia e Serviços de Saúde*, v. 24, n. 2, p. 325–332, jun 2015. ISSN 1679-4974. Citado 2 vezes nas páginas 16 e 32.

CHOO, J.; PARK, H. Customizing computational methods for visual analytics with big data. *IEEE Computer Graphics and Applications*, v. 33, n. 4, p. 22–28, 2013. ISSN 02721716. Citado na página 50.

CHOWKWANYUN, M. Big Data, Large-Scale Text Analysis, and Public Health Research. *American Journal of Public Health*, v. 109, n. S2, p. S126–S127, 2019. ISSN 0090-0036. Citado na página 44.

COAKLEY, M. F. et al. Unlocking the Power of Big Data at the National Institutes of Health. *Big Data*, v. 1, n. 3, p. 183–186, 2013. ISSN 2167-6461. Disponível em: <http://online.liebertpub.com/doi/abs/10.1089/big.2013.0012>. Citado na página 33.

COLLIER, K. et al. *A perspective on data mining*. [S.l.], 1998. Citado na página 52.

CORREIA, L. O. d. S.; PADILHA, B. M.; VASCONCELOS, S. M. L. Métodos para avaliar a completitude dos dados dos sistemas de informação em saúde do Brasil: uma revisão sistemática. *Ciência & Saúde Coletiva*, v. 19, n. 11, p. 4467–4478, nov 2014. ISSN 1413-8123. Disponível em: <http://www.scielo.br/scielo.php?script=sci{_}arttext{&}pid=S1413-81232014001104467{&}lng=p>. Citado na página 46.

CRAWFORD, K. The hidden biases in big data. *Harvard Business Review*, Cambridge, 2013. Disponível em: <https://hbr.org/2013/04/the-hidden-biases-in-big-data>. Citado na página 38.

De Mathias, T. A.; De Soboll, M. L. M. Confiabilidade de diagnósticos nos formulários de autorização de internação hospitalar. *Revista de Saude Publica*, v. 32, n. 6, p. 526–532, 1998. ISSN 00348910. Citado na página 46.

DEMCHENKO, Y. et al. Addressing big data issues in Scientific Data Infrastructure. *Proceedings of the 2013 International Conference on Collaboration Technologies and Systems, CTS 2013*, p. 48–55, 2013. Citado 3 vezes nas páginas 33, 34 e 35.

DICLEMENTE, R. et al. Need for Innovation in Public Health Research. *American journal of public health*, v. 109, n. S2, p. S117–S120, 2019. ISSN 15410048. Citado na página 16.

DONOHO, D. 50 Years of Data Science. *Journal of Computational and Graphical Statistics*, Taylor & Francis, v. 26, n. 4, p. 745–766, 2017. ISSN 15372715. Citado na página 14.

DOWDING, D. et al. Dashboards for improving patient care: Review of the literature. *International Journal of Medical Informatics*, Elsevier Ireland Ltd, v. 84, n. 2, p. 87–100, 2015. ISSN 18728243. Disponível em: <http://dx.doi.org/10.1016/j.ijmedinf.2014.10.001>. Citado na página 51.

ECKERSON, W. W. *Performance Dashboards: Measuring, Monitoring, and Managing Your Business*. New Jersey: John Wiley & Sons, 2006. 321 p. ISBN 9780471724179. Citado 2 vezes nas páginas 50 e 51.

EHRL, P. Minimum comparable areas for the period 1872-2010: an aggregation of Brazilian municipalities. *Estudos Econômicos (São Paulo)*, v. 47, n. 1, p. 215–229, 2017. Citado na página 47.

ENDERT, A.; BRADEL, L.; NORTH, C. Beyond control panels: Direct manipulation for visual analytics. *IEEE Computer Graphics and Applications*, v. 33, n. 4, p. 6–13, 2013. ISSN 02721716. Citado 2 vezes nas páginas 37 e 50.

FAN, C. et al. Automated mortality surveillance in South-Eastern Ontario for Pandemic Influenza Preparedness. *Canadian Journal of Public Health*, v. 101, n. 6, p. 459–63, 2010. Citado na página 51.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, v. 17, n. 3, 1996. ISSN 16113349. Citado 5 vezes nas páginas 36, 41, 42, 49 e 52.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. Knowledge discovery and data mining: towards a unifying framework. *Kdd-96*, p. 82–88, 1996. Citado 3 vezes nas páginas 41, 42 e 49.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, v. 39, n. 11, p. 27–34, nov 1996. ISSN 00010782. Disponível em: <http://portal.acm.org/citation.cfm?doid=240455.240464>. Citado na página 41.

FEW, S. *Information Dashboard Design*. O'Reilly, 2006. ISSN 00218499. ISBN 0596100167. Disponível em: <http://proquest.safaribooksonline.com/0596100167?suggested=top>. Citado na página 50.

FEW, S. *Information dashboard design: the effective visual communication of data*. New York: O'Reilly, 2006. 223 p. ISBN 0596100167. Citado 2 vezes nas páginas 50 e 51.

FORD, E. et al. Our data, our society, our health: A vision for inclusive and transparent health data science in the United Kingdom and beyond. *Learning Health Systems*, n. October 2018, p. 1–12, 2019. ISSN 23796146. Citado 2 vezes nas páginas 16 e 17.

FOTHERINGHAM, S.; ROGERSON, P. A. *The SAGE handbook of Spatial Analysis*. London: SAGE, 2009. ISBN 9781412910828. Citado na página 47.

FRANKE, B. et al. Statistical Inference, Learning and Models in Big Data. *International Statistical Review*, 2016. ISSN 03067734. Disponível em: <http://arxiv.org/abs/1509.02900http://doi.wiley.com/10.1111/insr.12176>. Citado na página 50.

FRAWLEY, W. J.; PIATETSKY-SHAPIRO, G.; MATHEUS, C. J. Knowledge Discovery in Databases: An Overview. *AI Magazine*, v. 13, n. 3, may 1992. Citado 2 vezes nas páginas 40 e 41.

FUNG, I. C.-H.; TSE, Z. T. H.; FU, K.-W. Converting Big Data into public health. *Science*, v. 347, n. 6222, p. 620–620, feb 2015. ISSN 0036-8075. Citado 2 vezes nas páginas 16 e 37.

GOTTGTROY, P. Ontology Driven Knowledge Discovery process: A proposal to integrate ontology engineering and KDD. In: *PACIS 2007 - 11th Pacific Asia Conference on Information Systems: Managing Diversity in Digital Enterprises*. [S.l.]: ACM Press, 2007. Citado 3 vezes nas páginas 52, 53 e 54.

HAN, J. et al. Survey on NoSQL database. In: *6th International Conference on Pervasive Computing and Applications*. Port Elizabeth, South Africa: IEEE, 2011. Citado na página 48.

HARRINGTON, L. et al. Nursing Research Dashboard Nursing Research Program. *Nurse Leader*, n. October 2006, 2006. Citado na página 51.

HAY, S. I. et al. Big Data Opportunities for Global Infectious Disease Surveillance. *PLoS Medicine*, v. 10, n. 4, p. e1001413, apr 2013. ISSN 1549-1676. Disponível em: <http://dx.plos.org/10.1371/journal.pmed.1001413>. Citado na página 16.

HAYASHI, C. What is data science? Fundamental Concepts and a Heuristic Example. In: HAYASHI, C. et al. (Ed.). *Data Science, Classification, and Related Methods*. Tokyo: Springer Japan, 1998. Citado na página 14.

HERLAND, M.; KHOSHGOFTAAR, T. M.; WALD, R. A review of data mining using big data in health informatics. *Journal Of Big Data*, v. 1, n. 1, p. 2, 2014. ISSN 2196-1115. Disponível em: <http://www.journalofbigdata.com/content/1/1/2>. Citado 4 vezes nas páginas 16, 33, 35 e 36.

HEY, T.; TANSLEY, S.; TOLLE, K. A transformed scientific method. In: *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond: Microsoft, 2009. p. xvii–xxxi. ISBN 978-0-9825442-0-4. Citado 2 vezes nas páginas 37 e 38.

HILTS, V. L. Aliis exterendum, or, the Origins of the Statistical Society of London. *Isis*, v. 69, n. 1, p. 21–43, mar 1978. ISSN 0021-1753. Disponível em: <https://www.journals.uchicago.edu/doi/10.1086/351931>. Citado 2 vezes nas páginas 15 e 16.

HOUSEH, M.; KUSHNIRUK, A. W.; BORYCKI, E. M. *Big Data, Big Challenges: A Healthcare Perspective*. Cham: Springer International Publishing, 2019. (Lecture Notes in Bioengineering). ISBN 978-3-030-06108-1. Citado 2 vezes nas páginas 14 e 16.

HUANG, T. et al. Promises and Challenges of Big Data Computing in Health Sciences. *Big Data Research*, Elsevier Inc., v. 2, n. 1, p. 2–11, 2015. ISSN 22145796. Disponível em: <http://dx.doi.org/10.1016/j.bdr.2015.02.002>. Citado 2 vezes nas páginas 33 e 37.

HUBER, P. J. *Data analysis: what can be learned from the past 50 years*. Hoboken: Wiley, 2011. ISBN 97811180106458. Citado na página 14.

IBM. *IBM Global Business Services Executive Report The value of analytics in healthcare*. Somers, 2012. 20 p. Citado na página 36.

JINPON, P.; JAROENSUTASINEE, M.; JAROENSUTASINEE, K. Business Intelligence and its Applications in the Public Healthcare System. *Walailak Journal*, v. 8, n. 1958, p. 97–110, 2011. Citado na página 51.

JORGE, M. H. P. d. M.; LAURENTI, R.; GOTLIEB, S. L. D. Análise da qualidade das estatísticas vitais brasileiras: a experiência do SIM e do SINASC. *Ciência e Saúde Coletiva2*, v. 12, n. 3, p. 643–654, 2007. Citado na página 45.

KELLING, S. et al. Data-intensive Science: A New Paradigm for Biodiversity Studies. *BioScience*, v. 59, n. 7, p. 613–620, 2009. ISSN 0006-3568. Citado na página 39.

KERZNER, H. *Project management: metrics, KPIs and dashboards*. 2. ed. New Jersey: John Wiley & Sons, 2013. v. 53. 1689–1699 p. ISBN 978111865895. Citado na página 50.

KHOURY, M. J.; IOANNIDIS, J. P. A. Big data meets public health. *Science*, v. 346, n. 6213, p. 1054–55, 2014. ISSN 0036-8075, 1095-9203. Citado 2 vezes nas páginas 14 e 36.

KITCHIN, R. Big data and human geography: Opportunities, challenges and risks. *Dialogues in Human Geography*, v. 3, n. 3, p. 262–267, 2013. ISSN 20438214. Citado 2 vezes nas páginas 34 e 38.

KITCHIN, R. Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, v. 1, n. 1, 2014. ISSN 2053-9517. Citado 4 vezes nas páginas 14, 33, 37 e 39.

KOSTKOVA, P. A roadmap to integrated digital public health surveillance. In: *Proceedings of the 22nd International Conference on World Wide Web - WWW '13 Companion*. New York, New York, USA: ACM Press, 2013. p. 687–694. ISBN 9781450320382. Disponível em: <http://www2013.org/companion/p687.pdfhttp://dl.acm.org/citation.cfm?doid=2487788.2488024>. Citado na página 51.

KOSTKOVA, P. et al. Integration and visualization public health dashboard. In: *Proceedings of the 23rd International Conference on World Wide Web - WWW '14 Companion*. New York, New York, USA: ACM Press, 2014. p. 657–662. ISBN 9781450327459. Disponível em: <http://dl.acm.org/citation.cfm?id=2567948.2579276{&}coll=DL{&}dl=GUIDE{&}CFID=574099975{&}CFTOKEN=12201186http://dl.acm.org/citation.cfm?doid=256>. Citado na página 51.

KUHN, T. S. *A Estrutura das Revoluções Científicas*. São Paulo: Perspectiva, 1970. Citado 2 vezes nas páginas 14 e 37.

LAZER, D. et al. The Parable of Google Flu: Traps in Big Data Analysis. *Science*, v. 343, n. 6176, p. 1203–1205, mar 2014. ISSN 0036-8075. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/24626916http://www.sciencemag.org/cgi/doi/10.1126/science.1248506>. Citado na página 36.

LIMA, C. R. d. A. et al. Revisão das dimensões de qualidade dos dados e métodos aplicados na avaliação dos sistemas de informação em saúde. *Cadernos de Saúde Pública*, v. 25, n. 10, p. 2095–2109, 2009. ISSN 0102-311X. Citado na página 46.

LIN, M.; LUCAS, H. C. Too Big to Fail : Large Samples and the p -Value Problem. *Information Systems Research*, v. 7047, p. 1–12, 2013. ISSN 1047-7047. Citado na página 50.

LOUKIDES, M. What is data science? *O'Reilly*, 2010. Citado na página 39.

MAASS, W. et al. Data-driven meets theory-driven research in the era of big data: Opportunities and challenges for information systems research. *Journal of the Association for Information Systems*, v. 19, n. 12, p. 1253–1273, 2018. ISSN 15583457. Citado na página 39.

MAITREY, S.; JHA, C. MapReduce: Simplified Data Analysis of Big Data. *Procedia Computer Science*, Elsevier Masson SAS, v. 57, p. 563–571, 2015. ISSN 18770509. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S1877050915019213>. Citado na página 49.

MARIN, H. d. F. Sistemas de informação em saúde: considerações gerais. *Journal of Heath Informatics*, v. 2, n. 1, p. 20–24, 2010. ISSN 2175-4411. Disponível em: <http://www.jhi-sbis.saude.ws/ojs-jhi/index.php/jhi-sbis/article/view/4/52>. Citado na página 43.

MILLER, H. J. The data avalanche is here. Shouldn't we be digging? *Journal of Regional Science*, v. 50, n. 1, p. 181–201, 2010. ISSN 00224146. Citado 2 vezes nas páginas 33 e 39.

Ministério da Saúde. *e-SUS*. 2019. Disponível em: <http://datasus.saude.gov.br/projetos/50-e-sus>. Citado na página 46.

MOONEY, S. J.; PEJAVER, V. Big Data in Public Health: Terminology, Machine Learning, and Privacy. *Annual Review of Public Health*, v. 39, n. 1, p. 95–112, apr 2018. ISSN 0163-7525. Disponível em: <http://www.annualreviews.org/doi/10.1146/annurev-publhealth-040617-014208>. Citado na página 16.

MOONEY, S. J.; WESTREICH, D. J.; EL-SAYED, A. M. Epidemiology in the Era of Big Data. *Epidemiology*, v. 26, n. 3, p. 390–394, may 2015. ISSN 1044-3983. Disponível em: <http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage{&}an=00001648-201505000-00>. Citado na página 16.

MURDOCH, T. B.; DETSKY, A. S. The Inevitable Application of Big Data to Health Care. *JAMA*, v. 309, n. 13, p. 1351, apr 2013. ISSN 0098-7484. Citado na página 36.

NUNES, E. D. Saúde Coletiva: história e paradigmas. *Interface - Comunic., Saude, Educ.*, v. 2, n. 3, p. 107–116, 1998. ISSN 1414-3283. Citado na página 24.

OLIVEIRA, M. M. de et al. Avaliação do Sistema de Informações sobre Nascidos Vivos. Brasil, 2006 a 2010. *Epidemiologia e Serviços de Saúde*, v. 24, n. 4, p. 629–640, oct 2015. ISSN 1679-4974. Citado na página 47.

PAHO. *Health system performance and improvement in the region of the Americas*. Washington: PAHO, 2011. 158 p. ISBN 9275073872. Citado na página 51.

PAIM, J. S.; De Almeida Filho, N. Saúde coletiva: Uma "nova saúde pública"ou campo aberto a novos paradigmas? *Revista de Saude Publica*, v. 32, n. 4, p. 299–316, 1998. ISSN 00348910. Citado na página 17.

PEDRAZA, D. F. Qualidade do Sistema de Informações sobre Nascidos Vivos (Sinasc): análise crítica da literatura. *Ciência e Saúde Coletiva*, v. 17, n. 10, p. 2729–2737, oct 2012. ISSN 1413-8123. Citado na página 47.

PETRUZALEK, D. *read.dbc: Read Data Stored in DBC (Compressed DBF) Files*. 2016. Disponível em: <https://cran.r-project.org/package=read.dbc>. Citado na página 48.

PIATETSKY-SHAPIRO, G. Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop. *AI Magazine*, v. 11, n. 4, 1990. Citado na página 40.

PIATETSKY-SHAPIRO, G. Knowledge Discovery in Databases: 10 years after. *SIGKDD Explorations*, v. 1, n. 2, p. 2–59, 2000. Citado na página 41.

PODGORNIK, M. N. et al. The Influenza Data Summary: A Prototype Application for Visualizing National Influenza Activity. In: *Intelligence and Security Informatics: Biosurveillance*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. p. 159–168. Citado na página 51.

PRESS, G. A Very Short History of Data Science. *Forbes*, may 2013. Disponível em: <https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/>. Citado na página 14.

PROVOST, F.; FAWCETT, T. Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, v. 1, n. 1, p. 51–59, 2013. ISSN 2167647X. Citado 2 vezes nas páginas 38 e 39.

PUTS, M.; DAAS, P.; WAAL, T. de. Finding errors in Big Data. *Significance*, v. 12, n. 3, p. 26–29, 2015. ISSN 17409713. Citado na página 50.

RANA, M. A. Building a Dashboard for the Punjab Health Department. *Asian Journal of Management Cases*, v. 12, n. 2, p. 128–147, sep 2015. ISSN 0972-8201. Disponível em: <http://www.scopus.com/inward/record.url?eid=2-s2.0-84943247309{&}partnerID=40{&}md5=dca42017993abe96eefb37619d18d8b0http://ajc.sagepub.com/cgi/doi/10.1177/0972820115>. Citado na página 51.

ROSEN, G. *A history of public health*. Baltimore: John Hopkins University Press, 1993. 441 p. ISBN 9781421416014. Citado 6 vezes nas páginas 19, 20, 22, 23, 24 e 25.

ROTONDO, A.; QUILLIGAN, F. Evolution Paths for Knowledge Discovery and Data Mining Process Models. *SN Computer Science*, Springer Singapore, v. 1, n. 2, p. 1–19, 2020. ISSN 2662-995X. Disponível em: <https://doi.org/10.1007/s42979-020-0117-6>. Citado 7 vezes nas páginas 40, 52, 53, 55, 56, 57 e 58.

SALATHÉ, M. et al. Digital Epidemiology. *PLoS Computational Biology*, v. 8, n. 7, p. e1002616, jul 2012. ISSN 1553-7358. Disponível em: <https://dx.plos.org/10.1371/journal.pcbi.1002616>. Citado na página 16.

SAS. *Managing the analytics life cycle for decisions at scale: how to go from data to decisions as quickly as possible*. [S.l.], 2016. Citado 2 vezes nas páginas 36 e 55.

SCHWARTZ, S.; SUSSER, E.; SUSSER, M. A Future for Epidemiology? *Annual Review of Public Health*, v. 20, n. 1, p. 15–33, may 1999. ISSN 0163-7525. Disponível em: <http://www.annualreviews.org/doi/10.1146/annurev.publhealth.20.1.15>. Citado na página 30.

SCLIAR, M. História do conceito de saúde. *Physis: Revista de Saúde Coletiva*, v. 17, n. 1, p. 29–41, apr 2007. ISSN 0103-7331. Citado na página 17.

SHMUELI, G. To Explain or to Predict? *Statistical Science*, v. 25, n. 3, p. 289–310, 2010. Citado 2 vezes nas páginas 15 e 16.

SIEGEL, E. *Predictive analysis*. Hoboken: John Wiley & Sons, 2013. Citado na página 38.

SUSSER, M.; STEIN, Z. *Eras in Epidemiology: The Evolution of Ideas*. New York: Oxford University Press, 2009. 368 p. ISSN 1098-6596. ISBN 9780199863754. Citado 8 vezes nas páginas 19, 20, 21, 22, 25, 26, 28 e 32.

SUSSER, M.; SUSSER, E. Choosing a future for epidemiology: I. Eras and paradigms. *American Journal of Public Health*, v. 86, n. 5, 1996. Citado 8 vezes nas páginas 19, 26, 28, 29, 30, 31, 32 e 37.

SUSSER, M.; SUSSER, E. Choosing a future for epidemiology: II. From black box to Chinese boxes and eco-epidemiology. *American Journal of Public Health*, v. 86, n. 5, p. 674–677, 1996. ISSN 00900036. Citado 4 vezes nas páginas 19, 30, 32 e 37.

TRUNK, A. An early concept of G.W. Leibniz regarding medicine. In: *Proceedings of the 2nd ICESHS*. Cracow: ICESHS, 2006. Citado na página 22.

TUKEY, J. W. The Future of Data Analysis. *The Annals of Mathematical Statistics*, v. 33, n. 1, p. 1–67, mar 1962. ISSN 0003-4851. Disponível em: <http://projecteuclid.org/euclid.aoms/1177704711>. Citado na página 15.

TUKEY, J. W. *Exploratory Data Analysis*. Reading: Addison-Wesley, 1977. ISBN 0201076160. Citado 2 vezes nas páginas 15 e 50.

TURNER, M. HealthStat: measuring the performance of the Irish Public Health Service. *Journal of the Statistical and Social Inquiry Society of Ireland2*, v. 38, 2009. Citado na página 51.

VAYENA, E. et al. Policy implications of big data in the health sector. *Bulletin of the World Health Organization*, v. 96, n. 1, p. 66–68, 2018. ISSN 15640604. Citado na página 16.

VIACAVA, F. Informações em saúde: a importância dos inqueritos populacionais. *Ciência e Saúde Coletiva*, v. 4, n. 7, p. 607–621, 2002. Citado na página 43.

VIACAVA, F. et al. Uma metodologia de avaliação do desempenho do sistema de saúde brasileiro. *Ciência & Saúde Coletiva*, v. 9, n. 3, p. 711–724, sep 2004. ISSN 1413-8123. Disponível em: <http://www.scielo.br/pdf/csc/v9n3/a16v09n3http://www.scielo.br/scielo.php?script=sci{_}arttext{&}pid=S1413-81232004000300021{&}lng=p>. Citado na página 51.

WALKER, T. C. The Perils of Paradigm Mentalities: Revisiting Kuhn, Lakatos, and Popper. *Perspectives on Politics*, v. 8, n. 2, p. 433–451, 2010. ISSN 1537-5927. Citado na página 38.

WANG, Y.; HAJLI, N. Exploring the path to big data analytics success in healthcare. *Journal of Business Research*, Elsevier Inc., 2016. ISSN 01482963. Disponível em: <http://dx.doi.org/10.1016/j.jbusres.2016.08.002>. Citado 2 vezes nas páginas 33 e 36.

WANG, Y.; KUNG, L.; BYRD, T. A. Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Social Change*, Elsevier Inc., p. 11, 2016. ISSN 0040-1625. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0040162516000500>. Citado na página 36.

WEIR, E. et al. Applying the balanced scorecard to local public health performance measurement: deliberations and decisions. *BMC public health*, v. 9, p. 127, 2009. ISSN 1471-2458. Citado na página 51.

YI, Q. et al. Integrating open-source technologies to build low-cost information systems for improved access to public health data. *International journal of health geographics*, v. 7, p. 29, 2008. ISSN 1476-072X. Citado na página 51.

Este documento foi digitado e diagramado utilizando as tecnologias TEX, LATEX, TEXLive e TEXMaker; com estilos providos pela classe ABNTEX2.

Gerado em 3 de abril de 2021.