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

TESE DE DOUTORADO

BRHIM - Base de Registros Hospitalares para Informações e Metadados

Tiago Andres Vaz
Orientadora: Profa. Dra. Suzi Alves Camey
Co-Orientador: Prof. Dr. Luís da Cunha Lamb

Porto Alegre, Junho de 2022

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

TESE DE DOUTORADO

BRHIM - Base de Registros Hospitalares para Informações e Metadados

Tiago Andres Vaz

Orientadora: Profa. Dra. Suzi Alves Camey
Co-Orientador: Prof. Dr. Luís da Cunha Lamb

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

CIP - Catalogação na Publicação

VAZ, TIAGO ANDRES
BRHIM - Base de Registros Hospitalares para Informações e Metadados / TIAGO ANDRES VAZ. -- 2022.
137 f.
Orientadora: SUZI CAMEY.

Coorientador: LUIS LAMB.

Tese (Doutorado) -- Universidade Federal do Rio Grande do Sul, Faculdade de Medicina, Programa de Pós-Graduação em Epidemiologia, Porto Alegre, BR-RS, 2022.

1. EPIDEMIOLOGIA. 2. INTELIGÊNCIA ARTIFICIAL. 3. ANONIMIZAÇÃO. 4. PRIVACIDADE. 5. APRENDIZADO DE MÁQUINA. I. CAMEY, SUZI, orient. II. LAMB, LUIS, coorient. III. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

Porto Alegre, Agosto de 2022

BANCA EXAMINADORA

Profa. Doutora Lisiane Pruinelli, Center for Nursing Informatics, University of Minnesota

Profa. Doutora Alessandra Dahmer, Programa de Pós-Graduação em Tecnologias da Informação e Gestão em Saúde, Universidade Federal de Ciências da Saúde de Porto Alegre

Prof. Doutor Ricardo de Souza Kuchenbecker, Programa de Pós-Graduação em Epidemiologia, Universidade Federal do Rio Grande do Sul

MENSAGEM

"Tente mover o mundo,
o primeiro passo será mover a si mesmo."
Platão

AGRADECIMENTOS

A minha esposa Gisele e aos meus filhos Vicente e Lucas,
obrigado por todo o apoio durante os últimos anos,
este trabalho não seria possível sem vocês.

Aos meus pais, Renato e Mana,
por terem me criado com amor, sempre ensinando o valor do conhecimento.

A Profa. Nadine Clausell e a Profa. Suzi Alves Camey,
por terem me incentivado e orientado ao longo desta jornada.

A Prof. Alessandra Drehmer, Prof. Airton Stein, Prof. Jesse Raffa, Prof. Luís Lamb, Prof. Mark Sendak, Prof. Miguel Dora, Enf. Ninon Girardon da Rosa, Prof. Ricardo Kuchenbecker, e ao Prof. Rodrigo Pires do Santos, gratidão por participarem da minha pós-graduação desde o começo.

Aos meus colegas do Hospital de Clínicas de Porto Alegre, que continuam usando e contribuindo para o crescimento e qualificação do AGHUse e dos projetos em que estive envolvido durante os últimos 20 anos.

A Fundação Médica do Rio Grande do Sul, por todas as parcerias nos cursos que realizamos sobre Ciência de Dados na Saúde.

Aos meus professores e colegas do Programa de Pós Graduação em Epidemiologia, pela grande amizade que criamos durante o aprendizado.

Dedico este trabalho em memória do
Prof. Amarilio de Macedo Vieira da Cunha.

SUMÁRIO

ABREVIATURAS E SIGLAS

AGHUse - Aplicativos de Gestão Hospitalar

BRHIM - Base de Registros Hospitalares para Informações e Metadados

DP - Desvio Padrão

HDL - Hight density lipoprotein (lipoproteína de alta densidade)

HIPAA - *Health Insurance Portability and Accountability Act* (Ato de portabilidade e responsabilidade dos planos de saúde)

GBD - *Global Burden of Diseases* (Visão Global das Doenças)

GPL - *General Public License* (Licença Pública Geral)

GDPR - *General Data Protection Regulation* (Regulação Geral para Proteção de Dados)

I2B2 - *Informatics for Integrating Biology and the Bedside* (Informática para integração da Biologia e a beira do leito)

IA - Inteligência Artificial

LDL - *Low density lipoprotein* (lipoproteína de baixa densidade)

MIMIC - *Medical Information Mart for Intensive Care* (Banco de Dados Médico para Cuidados Intensivos)

NHS - *National Health Services* (Sistema Nacional de Saúde da Inglaterra)

HIS - Hospital Information System

EHR - Electronic Health Records

EMR - Electronic Medical Records

LGPD - Lei Geral de Proteção de Dados Pessoais

ORHBR - Ontologia para Anonimização de Registros Hospitalares no Brasil

PHI - Protected Health Information (Informação de Saúde Protegida)

S-RES - Sistemas de Registros Eletrônicos de Saúde

SUS - Sistema Único de Saúde

TI - Tecnologia da Informação

VisTA - *Veterans Health Information Systems and Technology Architecture* (Sistema e Arquitetura de Tecnologia da Saúde dos Veteranos)

RESUMO

Os riscos de reidentificação de dados hospitalares são altos e há uma demanda por eles em projetos de desenvolvimento e validação de Inteligência Artificial (IA). Este trabalho aborda os principais métodos de preparação de registros hospitalares usados para realizar estudos observacionais de maneira direcionada de avaliar o risco de reidentificação e o impacto da perda de informações que a anonimização produz nos resultados da IA. Uma revisão sobre o assunto é apresentada no início e após são apresentados dois artigos, sempre considerando o contexto da utilização de registros hospitalares em estudos epidemiológicos. O primeiro artigo propõe uma ontologia de domínio para definir um escopo para a tratar a anonimização. Os tipos de ataques, os tipos de dados e atributos, os modelos de privacidade, os tipos de uso da inteligência artificial e os diferentes delineamentos são apresentados. Foi feito um exemplo de instância da ontologia na ferramenta Web Protegé, disponível pela Universidade de Stanford para a construção de ontologias e que permite replica-la. O segundo artigo visa definir uma receita de preparação de prontuário hospitalar com 5 etapas para implementar a pseudo-anonimização, desidentificação e anonimização de dados e comparar os efeitos dessas etapas em uma aplicação da IA. Para isto, um evento Datathon foi realizado para desenvolver um preditor de IA de mortalidade hospitalar. Comparando os resultados da IA usando os dados originais e os dados anônimos, demonstrando uma diferenca inferior a 1% nos resultados da AUC-ROC, enquanto o risco de um paciente ser identificado foi reduzido em 95%, demonstrando que o preparo pode ser sistematizado agregando privacidade e computando a perda de informações, a fim de torná-los transparentes.

ABSTRACT

The risks of re-identifying hospital data is high and there is a demand for them in projects for the development and validation of Artificial Intelligence (AI). This approach addresses the main methods of preparing hospital records used to carry out observational studies and in a directed way to assess the risk of re-identification and the impact of the loss of information that anonymization produces on AI results. A review of the review on the subject is presented at the beginning and after the literature is presented two articles, always considering the context of the use of hospital records in epidemiological studies. The first article proposes a domain ontology to define a scope for the search for anonymization. The types of attacks, the types of attacks, the types of data and attributes, the privacy models, the types of use that artificial intelligence devices and the different delineations are presented. An example of an ontology instance was made in the Web Protegé tool, made available by Stanford University for building ontologies and which allows replicating pregnant children and thus disseminating anonymization atology. The article aims to define a second hospital record preparation recipe with 5 steps for implementing pseudo-anonymization, de-identification and data anonymization and to compare the effects of these steps in an AI application. A Datathon event was conducted to develop an AI predictor of hospital mortality. Comparing the AI results using the original data and the anonymized data, which were identified as less than 1% results on the AUC-ROC, while the risk of a registered patient was recorded at 95%, demonstrating that the preparation can be systematized with privacy privacy and information loss in order to make them transparent.

1. APRESENTAÇÃO

Este trabalho consiste na tese de doutorado intitulada “BRHIM - Base de Registros Hospitalares para Informações e Metadados”, a ser apresentada ao Programa de Pós-Graduação em Epidemiologia da Universidade Federal do Rio Grande do Sul, no segundo semestre de 2022.

Aluno: Tiago Andres Vaz

Orientadora: Profa. Dra. Suzi Alves Camey

Co-Orientador: Prof. Dr .Luís da Cunha Lamb

Colaborador: Prof. Dr. José Miguel Dora

O trabalho é apresentado em três partes:

1. Introdução e Revisão da Literatura.
2. Objetivos.
3. Artigos.

As referências foram organizadas e formatadas utilizando o gerenciador de referências Paperpile.

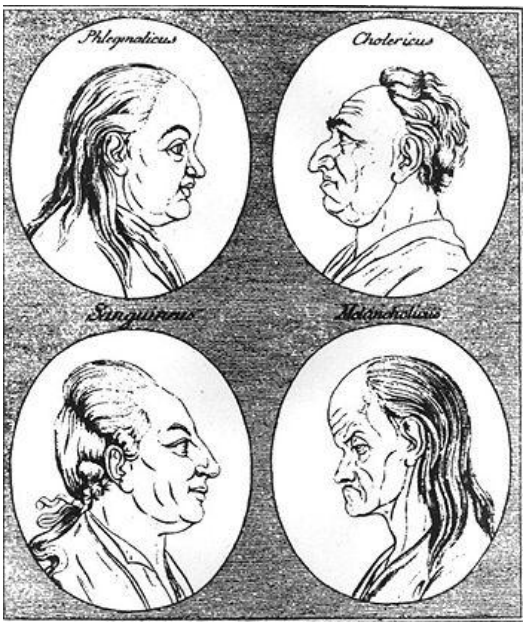
2. INTRODUÇÃO

Na perspectiva teórico e conceitual a epidemiologia iniciou o estudo de doenças com o uso de dados no final do século XIX, utilizando análises estatísticas, comparando subgrupos populacionais e comunicando os resultados, que com o tempo passaram a ser preservados, compartilhados e compreendidos pelos epidemiologistas de diferentes formas (CAMERON; JONES, 1983; KRIEGER, 2011; RILEY, 2001).

Neste capítulo inicial faremos uma retrospectiva, importante para entendermos o tema deste trabalho. Antes da epidemiologia, os estudos centrados na análise de dados relacionados aos problemas de saúde foram feitos de forma restrita, com o objetivo de controlar doenças transmissíveis e para compreender as relações entre as condições do ambiente e doenças específicas. Estes estudos evoluíram em seus métodos de pesquisa, agregaram novas tecnologias e com o tempo forneceram dados que suportaram a formulação de diferentes teorias epidemiológicas. (ALMEIDA FILHO, 1986).

Na Grécia antiga de dois mil e quatrocentos anos atrás, Hipócrates iniciou o registro de dados sobre a saúde das pessoas para servir como um artefato na transmissão do conhecimento para seus alunos. Suas análises o levaram ao desenvolvimento de conceitos fundamentais para a medicina, entre eles a observação e descrição de diferentes tipos de informações, para após realizar a classificação de cada paciente em um dos quatro tipos de Humores definidos por ele (Ver Figura 1). Isto era feito agrupando dados de sinais e de sintomas dos pacientes, dando a oportunidade para a visualização de conjuntos populacionais e o estabelecimento de uma compreensão auto-evidente na relação existente entre os dados coletados. Analisando dados, Hipócrates fundou a medicina (GRAMMATICOS; DIAMANTIS, 2008; KLEISIARIS; SFAKIANAKIS; PAPATHANASIOU, 2014).

Figura 1: Os quatro humores de Hipócrates



Fonte: LAVATER (1969)

Passaram dois mil anos e os registros de saúde evoluíram timidamente em bibliotecas privadas de acesso ao público, onde novas teorias surgiram e sucumbiram ao redor do mundo. Foi somente em 1543 na europa renascentista que Andreas Vesalius publicou “De humani corporis fabrica”, descrevendo a anatomia do corpo em detalhes e trazendo notoriedade para uma série de novas revelações sobre a nossa espécie, incluindo tamanho e peso dos nossos órgãos, a quantidade de ossos, os tipos de tecidos, entre outras características. Nestes moldes, muitos avanços seguiram acontecendo durante a renascença, criando uma disruptura com a medicina hipocrática e dando início a uma nova era de luz através da ciência (MARGÓCSY; SOMOS; JOFFE, 2019; SIRAISI, 2009).

Logo após, em 1662 o inglês John Graunt publicou uma análise sobre a mortalidade em Londres, estudando quantitativamente a natalidade e a incidência de doenças,

considerando as diferenças encontradas em grupos de homens e de mulheres. Em 1800 Willian Farr ainda dava seguimento as ideias quantitativas de Graunt, aprimorando os estudos estatísticos e iniciando a comunicação de alertas de saúde para as autoridades e também para a população. Mas isto não foi o suficientes para alterar a realidade estabelecida pela teoria dos miasmas, ainda dos tempos hipocráticos, que propunha minimizar os problemas de saúde combatendo a insalubridade com a eliminação de odores e da "poluição" provocada por supostas partículas existentes no ar. A teoria dos miasmas foi utilizada como referência para combater a cólera pelos governantes na Inglaterra entre 1831 a 1866 e deixou mais de 52 mil mortos em Londres (HALLIDAY, 2001; HUMPHEYS, 1885; SUTHERLAND, 1963).

Mas foi esta epidemia de cólera que levou um renomado médico anestesista chamado John Snow, notório por ter recebido a determinação de realizar o parto da rainha sem dores, a conduzir um experimento inovador para a época e dando grande visibilidade para os os dados analisados por ele (PANETH, 2004; ZUCK, 2004).

Através da compreensão quantitativa e qualitativa elementar das informações, Snow desvendou um novo método para encontrar as variáveis determinantes na ocorrência de uma doença. Utilizou dados para tentar combater a epidemia de cólera, comparando a incidência de casos, em diferentes grupos de pessoas que beberem a água proveniente de fontes distintas de abastecimento no centro de Londres (CAMERON; JONES, 1983).

Desta forma, a sociedade científica acompanhou os resultados do passo gigantesco dado por John Snow e após a sua morte em 1858, temos o surgimento da epidemiologia moderna (HALLIDAY, 2001).

Sucederam as suas ideias, teorias relacionadas aos germes, aos gens, ao reducionismo - que é uma perspectiva mecanicista para explicação dos problemas de saúde, todas elas advindas do pensamento epidemiológico. De forma aplicada, a epidemiologia não só

enfrentou a cólera, mas também erradicou a varíola e hoje é aceita mundialmente, sendo aplicada em comunidades de todos os portes (HENDERSON, 1972).

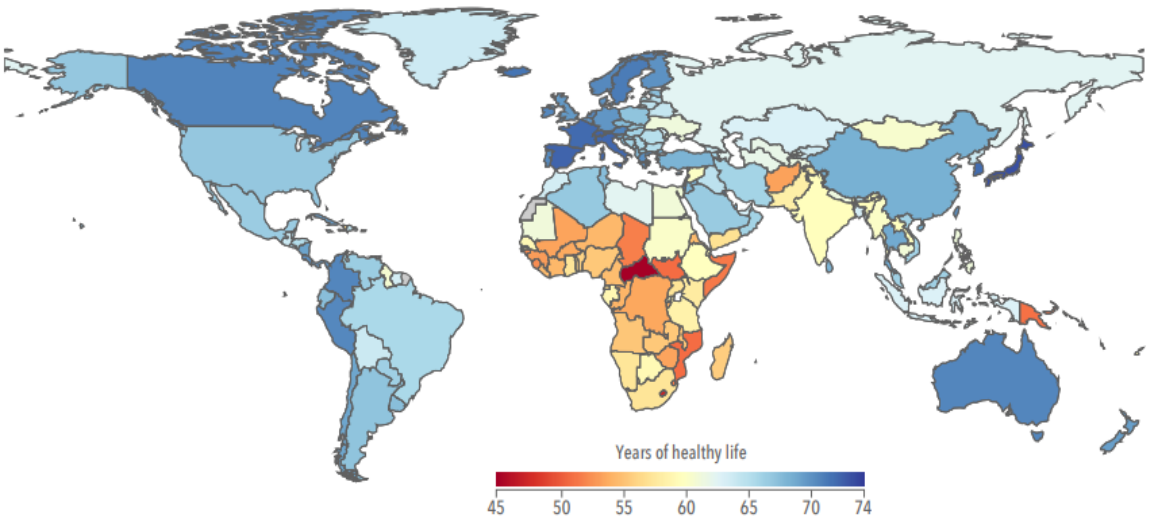
Uma das principais contribuições resultadas da evolução das teorias na epidemiologia e que merece um destaque é o aumento da expectativa de vida da população global ano após ano. Entre 1800 e o ano 2000 a média da expectativa global de vida foi de 30 anos para 67 anos (RILEY, 2001). De acordo com o estudo Global Burden of Diseases a expectativa de vida global em 2017 é de 71 anos para os homens e 76 anos para as mulheres (INSTITUTE FOR HEALTH METRICS AND EVALUATION (IHME), 2018).

Faz parte do raciocínio epidemiológico explorar novas tecnologias e avançar em campos onde ainda temos resultados insuficientes. Com a crescente informatização da saúde, principalmente nos grandes centros hospitalares, surgiram dados que vão além do tradicional escopo dos dados governamentais e hoje são utilizados em estudos como o Global Burden of Diseases (Ver Figura 2). Estas bases de dados podem conter registros agregados ou individuais e detalhados sobre pessoas naturais, onde pesquisadores podem testar suas hipóteses com novos métodos de análise e descoberta de informações, incluindo as técnicas de Aprendizado de Máquina e o uso de Inteligência Artificial (DE STATISTIQUE APPLIQUÉE; CEDEX, *[s. d.]*; FLOURIS; DUFFY, 2006).

Analisar dados é preciso. A partir daqui, veremos as teorias que fundamentam o uso de dados em estudos epidemiológicos, descreveremos os tipos de dados existentes, passando pelas grandes bases hospitalares, até chegar nas bases de dados governamentais utilizadas para pesquisa em saúde. Encerraremos aprofundando o conhecimento sobre as técnicas existentes para fornecer dados anonimizados e como elas podem ser utilizadas para o treinamento de sistemas com Inteligência Artificial, explorando o contexto científico recente desta disciplina da Ciência da Computação e que tem cada vez mais chamado a atenção dos

epidemiologistas, na medida em que se percebe o surgimento de uma ferramenta que demandará novas formas de pensar os estudos centrados na análise de dados da saúde (FLOURIS; DUFFY, 2006; LAVIGNE *et al.*, 2019; RAJKOMAR *et al.*, 2018; SHABAN-NEJAD; MICHALOWSKI; BUCKERIDGE, 2018).

Figura 2: Idade que alguém pode esperar viver com saúde plena.



Exemplo de Análise de Dados na Saúde, que mostra a iniquidade entre os países na expectativa de vida ao nascer considerando a mortalidade e a invalidez para ambos os sexos em 2017. Fonte: Publicado originalmente em (INSTITUTE FOR HEALTH METRICS AND EVALUATION (IHME), 2018)

3. REVISÃO DE LITERATURA

3.1. TIPOS DE DADOS

Dados são observações de fatos e ideias, ou resultados de medição que podem ser representados formalmente e documentados fisicamente de diferentes formas, pelas quais a partir do seu estado original é possível obter informações. Dados digitais são aqueles que foram traduzidos para linguagem computacional, permitindo assim processamento e a comunicação de forma otimizada. Os tipos de dados digitais existentes podem ser classificados em diferentes dimensões (Ver Quadro 1) que auxiliam a estabelecer uma ontologia de domínio, na compreensão dos recursos necessários e na utilidade do seu processamento (CHECKLAND; HOLWELL, 1998; KEET *et al.*, 2015; KLÖSGEN, 2002; PANOV; SOLDATOVA; DŽEROSKI, 2016).

3.1.1 Dimensões para Classificação

Tradicionalmente em estudos epidemiológicos, classificamos os tipos de dados em variáveis de acordo com a sua natureza: qualitativas ou quantitativas. Qualitativas são as variáveis que representam classes de informação, podendo ser categóricas nominais ou ordinais (quando existe ordenação entre as diferentes categorias). As variáveis Quantitativas são aquelas mensuráveis com valores em uma escala numérica, podendo assumir valores inteiros (variáveis discretas) ou valores dentro de uma escala (variáveis contínuas). Algumas vezes, um tipo de dado pode ser transformado de acordo com a necessidade do estudo realizado, seja agrupando informação (por exemplo: cor da pele classificada em branca e não branca), ou realizando operações matemáticas como logaritmo e raiz quadrada. (DUNN; CLARK, 2009; HAI DATA AND STATISTICS | HAI | CDC, *[s. d.]*; LENZ, 2009).

Quando vamos criar programas de computador, ao escrever novos código de programação surge a necessidade de classificar os dados na perspectiva da Ciência da Computação, onde um tipo de dado é considerado um atributo pelo qual compiladores e interpretadores orientam a sua execução. A quantidade de classificações existentes para definir estes tipos de dados é proporcional à quantidade de linguagens de programação existentes, tornando impraticável esgotar o tema. Um exemplo clássico da literatura técnica em informática, apresenta os tipos de dados na linguagem C++: int, float, char e bool, respectivamente utilizados para representar números inteiros, números reais, caracteres e booleanos. As máquinas também podem trabalhar com tipos de sistemas numéricos diferentes do decimal, como no exemplo da paleta de cores HTML que apresenta dados representados no formato hexadecimal (i.e.: 0A2HFF) e no caso dos dados binários, que utilizam geralmente 0/1 ou falso/verdadeiro, positivo/negativo para representar valores booleanos. O sistemas de informação também colocam a perspectiva do negócio para classificar diferentes tipos de dados de acordo com o formato (áudio, vídeo, texto, e outros) e o setor de atividade, como são os dados da Saúde, Segurança, Financeiros, Recursos Humanos e outras necessidades empresariais, usualmente consolidadas no mercado em bases de dados do tipo *Data Warehouse*. A grande quantidade de informação existente gerou a necessidade de classificar os dados de formas ainda mais distintas, agora realizadas com a colaboração dos usuários utilizando marcadores taxonômicos que dão origem aos metadados, sendo o tipo “tag” ou “label” o metadado mais conhecido. Metadados são dados que ajudam a descrever os dados em questão. (DALE; WALKER, 1992; KELLEY; POHL, 1994; KIMBALL *et al.*, 1998; KLÖSGEN, 2002).

Quadro 1: Tipos de dados digitais existentes classificados de acordo com a dimensão

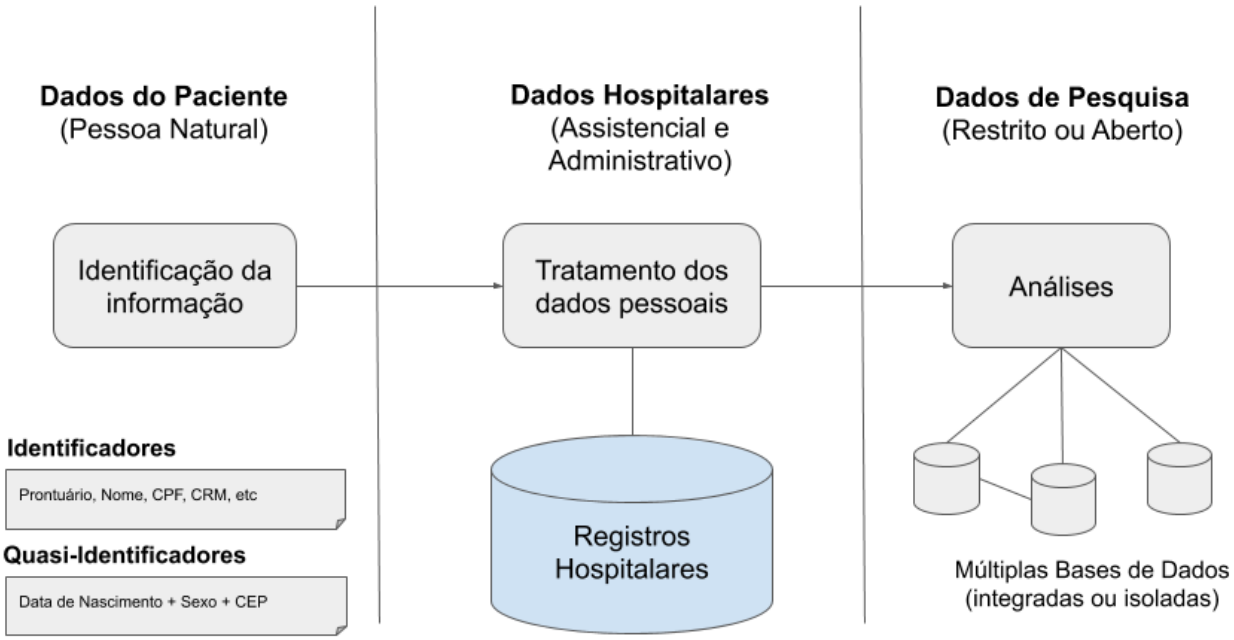
Nro	**Dimensões**	**Exemplos de classificação dos dados**	**Fonte**
1	Formato	Código, Texto, Imagem, Multimídia, Hiperlink	(DALE; WALKER, 1992)
2	Origem	Hospitalar, Bancário, Jurídico e outros	(KIMBALL *et al.*, 1998)
3	Natureza	Categórico (Ordinal e Nominal) e Numérico(Discreto e Contínuo)	(LENZ, 2009)
4	Completude	Completo, Incompleto (dados faltantes)	(KLÖSGEN, 2002)
5	Metadados	Dicionário, Classes de Domínio, Marcadores	(KLÖSGEN, 2002)
6	Sistema de numeração	Binário, Decimal, Hexadecimal	(TANNA, 1 julho 2018)
7	Estrutura	Estruturado (registro, tabela, conjunto), Não Estruturado (texto, imagem, som), Semi Estruturado (JSON, XML)	(CHEN, 1976), (KLÖSGEN, 2002) (BUNEMAN, 1997)
8	Conjunto	Numérico (natural, inteiro, fracionário, irracional, real), Literal (letras e símbolos), Lógico (verdadeiro, falso)	(KLÖSGEN, 2002)
9	Espacialidade	Pontos, Linhas, Áreas, Superfícies, Distâncias	(KLÖSGEN, 2002)
10	Agregação e Perturbação	Micro-Dados, Agregados, Imputados, Ruídos	(KLÖSGEN, 2002)
11	Tipagem	Caracter (ASCII, UTF, ISO), inteiro, ponto flutuante de precisão simples, ponto flutuante de dupla precisão e lógico	(CHARACTER ENCODINGS: ESSENTIAL CONCEPTS, *[s. d.]*) (KLÖSGEN, 2002)
12	Privacidade	Pessoal, Pessoal Sensível, Restrito, Aberto, Pseudonimizado, Anonimizado	(BRASIL, 2018)
13	Localização	Português-BR, Inglês-EUA e outras	(KLÖSGEN, 2002)

Fonte: Elaborado pelos autores (2022).

Para aprofundar a compreensão dos tipos de dados existentes nos grandes bancos de dados de registros hospitalares, estabelecemos uma ontologia de domínio para organizar os tipos de dados hospitalares quanto a sua perspectiva de privacidade, dividida em sub-tópicos designados conforme a ordem estabelecida na captura e no uso destes dados. Colocando foco no paciente e ajudando a compreender como acontece o processamento dos seus dados

pessoais ao longo do atendimento em um hospital (ver figura 3) (QUEIROZ; LINO; GUSTAVO H M, 2016).

Figura 3: Contexto de uma Base de Registros Hospitalar. Iniciaremos entendendo o que



Fonte: Imagem elaborada pelo autor.

são **dados pessoais** de acordo a Lei Geral de Proteção de Dados Pessoais em vigor no Brasil e como eles preservam os atributos que permitem a identificação de um paciente e dos profissionais envolvidos ao longo da sua assistência. A partir desta compreensão, analisaremos os **dados hospitalares** que são gerados em um sistema de gestão hospitalar informatizado. Por fim, descreveremos os **dados de pesquisa** e as bases governamentais de saúde onde parte destes dados são informados e processados, utilizando exemplos existentes no Brasil e nos Estados Unidos. Compreendendo o ciclo de vida dos dados na saúde, vamos

entender a utilidade destas bases de dados para epidemiologia e para definição de políticas públicas de saúde (BRASIL, 2018; SULLIVAN, 2004).

3.1.2 Dados do Paciente

Os dados que podem identificar um paciente são aqueles que podem comprometer a sua privacidade, um direito constitucional no Brasil (Inciso X do Artigo 5 da Constituição Federal de 1988) e em outros 130 países. A privacidade é considerada um direito fundamental pela ONU, que a define como algo essencial para proteção da dignidade humana e que forma a base para o estabelecimento dos direitos humanos internacionalmente. (ART. 5, INC. X DA CONSTITUIÇÃO FEDERAL DE 88, *[s. d.]*; BOWIE; JAMAL, 2006; SOLOVE, 2008).

Ela surge para permitir que os indivíduos definam seus próprios limites no estabelecimento de suas relações, evitando que interferências e agressões desproporcionais ocorram em suas vidas. Institutos internacionais que observam as questões de privacidade apontam que os riscos da quebra de sigilo por definição, podem colocar em risco o seu corpo, os seus bens, a sua família, os grupos e as pessoas com quem você convive, a sua comunicação (incluindo telefone e internet) e toda sua correspondência (digital e analógica) (PARKER, 2017).

A Lei Geral de Proteção de Dados Pessoais (LGPD) do Brasil foi criada em 2018 seguindo os passos iniciados em 2016 pela União Europeia com a General Data Protection Regulation (GDPR). Estas leis impõem regulamentação para o controle e penalidades para casos em que a legislação não é respeitada. São focadas na segurança e na privacidade com que as organizações lidam com **dados pessoais**, ou seja, dados identificados ou identificáveis que pertencem a pessoas naturais, ou seja, pertencem e identificam quem são os indivíduos

titulares dos dados. Quando os dados não podem mais ser identificados, eles deixam de ser dados pessoais e não estão mais sujeitos às regulamentações. (BRASIL, 2018; DENLEY; FOULSHAM; HITCHEN, 2019; SULLIVAN, 2004).

A LGPD do Brasil, a exemplo da GDPR Européia, define **dado pessoal** como toda informação relacionada a pessoa natural identificada ou identificável. O chamado **dado pessoal sensível** é definido como todo o dado pessoal sobre origem racial ou étnica, convicção religiosa, opinião política, filiação a sindicato ou a organização de caráter religioso, filosófico ou político, dado referente à saúde ou à vida sexual, dado genético ou biométrico, quando vinculado a uma pessoa natural. Todos os dados identificáveis sobre a saúde das pessoas são considerados um dado pessoal sensível e devem ser tratados considerando a LGPD (BRASIL, 2018).

Sem a privacidade, a sociedade estaria a mercê do uso arbitrário e injustificado do poder, com risco de ter sua honra e reputação prejudicada, privada de liberdade para pensar e orientada a discriminação. Os aspectos da privacidade, garantem aos indivíduos o direito de ficarem sozinhos, a estabelecerem a sua intimidade e embasar o comportamento social histórico de "fechar a porta de casa". (DENLEY; FOULSHAM; HITCHEN, 2019; WHAT IS PRIVACY?, *[s. d.]*).

De acordo com a organização sem fins lucrativos inglesa Privacy International, estamos vivendo um momento em que a definição de ética da vida moderna, cercada por regras de conduta orientadas ao comércio e a mercê do poder do estado, ainda está sendo definida. Isto está acontecendo na medida em que surgem novos algoritmos que identificam pessoas utilizando dados provenientes de todos os lugares da internet na tentativa de corromper a privacidade. O conceito do acesso limitado a informação das pessoas estabelecido na GDPR e na LGPD, deve permitir que cada um controle quem, quando e onde

nossas informações podem ser utilizadas, e por definição, dados sensíveis devem ser secretos e as pessoas devem ter a opção de revelá-los ou não (BENNETT; RAAB, 2017; WHAT IS PRIVACY?, *[s. d.]*).

O conceito de invasão de privacidade pode variar entre diferentes culturas, e se sobrepõem parcialmente ao conceito de segurança, pois ambos tratam sobre a definição dos termos de uso dos dados e também sobre a proteção da informação, visto que um dos principais riscos à privacidade das pessoas é a falta de proteção dos dados eletrônicos pessoais. Danos deste tipo, geralmente não são percebidos no momento em que acontecem e muitas vezes não deixam rastros do local e detalhes sobre como ocorreu a exposição. Estes vazamentos tão pouco alertam as pessoas que foram afetadas, potencializando o roubo de identidade e tornando elas indefesas ou submetidas a um prejuízo, que pode ser irreversível. Preocupada com este tipo de dano que tem potencial massivo de atingir as pessoas, organizações internacionais como a Organização para Cooperação do Desenvolvimento Econômico (OCDE) publicam guias de proteção e privacidade na expectativa de que os países integrantes destas organizações assumam o compromisso de evitar que órgãos públicos, partidos e grandes corporações abusem da privacidade dos indivíduos internacionalmente (LEFEVRE; DEWITT; RAMAKRISHNAN, 2008; OECD; OECD, 2003; QUEIROZ; LINO; GUSTAVO H M, 2016).

Os atributos que permitem identificar diretamente uma pessoa geralmente são o **nome completo** e os Números de Documentos (passaporte, código do prontuário, carteira nacional de habilitação, registro de identidade, conselhos profissionais, entre outros). Mas uma pessoa também é identificável por outros dados pessoais chamados de quasi-identificadores, que somados podem revelar a identidade do titular. A pesquisadora L. Sweeney em 2015 conduziu um experimento utilizando quasi-identificadores e re-identificando pessoas em uma

série de documentos públicos nos EUA considerados anônimos até então, dando início a uma nova área de pesquisa relacionada a reidentificação de bases de dados (SWEENEY, 2015).

Na área da saúde os Estados Unidos regulou o uso de dados em 1996 através da *Health Insurance Portability and Accountability Act* (HIPAA). Para fins de abertura dos dados na saúde, a desidentificação necessária para o compartilhamento de dados na saúde é obtida removendo 18 itens considerados dados protegidos e chamados de *Protected Health Information* (PHI), incluindo: nomes, endereços (incluindo código postal), todas as datas e todas as informações de contato (e-mail, telefone) e fotos que possam identificar as pessoas (SNELL, 2017) (Ver Quadro 2). Este é um processo considerado balanceado pelas autoridades nos EUA, pois fornece privacidade (entendendo os riscos de que é possível re-identificar as pessoas), viabilizando os negócios digitais entre os principais envolvidos na assistência à saúde. Fornecedores de serviço de nuvem para hospitais, a exemplo da Amazon, Microsoft e Google adaptaram seus serviços que atendem as demandas da HIPAA, entretanto não existe uma certificação formal para avaliar a qualidade e a segurança destas implementações. As empresas oferecem criptografia do lado do servidor e do cliente e vários métodos de gerenciamento de chaves. Existem muitos detalhes técnicos demandados para implementar corretamente estas soluções em cada caso de uso, por exemplo: as conexões entre servidores da nuvem que possam conter PHI devem utilizar transporte de dados criptografado, demandando diferentes configurações customizadas para o perfil de cada hospital. Por isto, cabe aos hospitais utilizarem estes serviços de forma correta, capacitando seu quadro funcional para evitar usar PHI em nomes de diretório e arquivos, em dados de testes, em objetos de programação ou em metadados de configuração do *software*.(BRACCI; CORRADI; FOSCHINI, 2012; MALIN; BENITEZ; MASYS, 2011; SULLIVAN, 2004; SZARVAS; FARKAS; BUSA-FEKETE, 2007)

Quadro 2: Dados que demandam tratamento para privacidade nos EUA

Nro	Informação Pessoal de Saúde (PHI)
1	Nome
2	Endereço (todas as subdivisões geográficas menores que o estado, incluindo endereço, município e CEP)
3	Todos os elementos (exceto anos) de datas relacionadas a um indivíduo (incluindo data de nascimento, data de admissão, data de alta, data de falecimento e idade exata se tiver mais de 89 anos)
4	Números de telefone
5	Número de fax
6	Endereço de e-mail
7	Número da Segurança Social
8	Número do prontuário médico
9	Número do beneficiário do plano de saúde
10	Número da conta
11	Número do certificado ou licença
12	Identificadores de veículos e números de série, incluindo números de matrículas
13	Identificadores de dispositivo e números de série
14	URL da Web
15	Endereço IP (Internet Protocol)
16	Identificadores biométricos: digital de dedo, retina ou voz
17	Imagem fotográfica - As imagens fotográficas não se limitam às imagens do rosto.
18	Qualquer outra característica que possa identificar exclusivamente o indivíduo: número, característica ou código de identificação exclusivo, exceto o código exclusivo atribuído pelo investigador para codificar os dados

Fonte: SULLIVAN (2004)

3.1.3 Registros Eletrônicos Hospitalares

De acordo com a LGPD o tratamento de dados consiste em toda operação realizada com dados pessoais, incluindo a coleta, armazenamento, produção, eliminação,

processamento, comunicação ou transferência de dados, entre outros. Em resumo, vamos dizer que após a coleta dos dados de identificação do paciente, são capturados pelos sistemas de informação nos hospitais durante os processos de trabalho assistencial dois grandes conjuntos de dados: os dados transacionais e os dados observacionais (BRASIL, 2018).

Os **dados transacionais** são decorrentes de todas as transações gerenciadas pelos sistemas de gerenciamento de banco de dados (SGBD), que vão guardando a história transação por transação (quem, quando, onde) das operações de consulta, edição, criação e exclusão de dados. Inclui todos os dados que são gerados automaticamente durante os fluxos de execução dos processos hospitalares, incluindo os cálculos e todos os *logs* de sistema e das funcionalidades assistenciais e administrativas que foram utilizadas. Os sistemas transacionais exigem usuário e senha dos funcionários, que passam por variadas etapas de concessão de permissões de acesso antes de terem autorização para realizarem transações no sistema (HRIPCSAK *et al.*, 2015; TIDKE; MEHTA; DHANANI, 2018; TROVATI *et al.*, 2016).

Já os **dados observacionais** são coletados de diferentes formas para descrever o pacientes e os seus eventos. São registros manuais feitos pelos profissionais da saúde e pela equipe administrativa, tanto em texto livre quanto através de formulários estruturados, ou feitos por máquinas que realizam testes de diagnóstico (gráfico, som, imagem) ou monitoram sinais e sintomas dos pacientes, medindo, calculando e coletando observações. Novas máquinas capazes de coletar observações surgem na medida que a automação e a internet das coisas ganham popularidade, mas a interoperabilidade destes recursos com os sistemas transacionais descritos anteriormente é um problema que ainda demanda um alto investimento em um setor extremamente complexo para implementar padronizações em larga escala (GARZA *et al.*, 2016; KIMBALL *et al.*, 1998).

Nos hospitais, o **tratamento dos dados** gerados pelos sistemas transacionais e por todas as demais observações é realizado no âmbito da Tecnologia da Informação e Comunicação (TIC), mas envolve a maior parte dos funcionários do hospital. O principal termo utilizado no Brasil para designar os sistemas de TIC utilizados para gestão clínica e administrativa nos hospitais é Sistemas de Registros Eletrônicos de Saúde (S-RES). Popularmente, o S-RES também pode ser conhecido pelos termos "Prontuário Eletrônico", "Sistema de Gestão Hospitalar" ou ainda pelas siglas em inglês: HIS (Hospital Information System), EHR (Electronic Health Records) e EMR (Electronic Medical Records). (SOUZA *et al.*, 2019).

Os hospitais adotam soluções de S-RES fornecidas pelo mercado, ou desenvolvem seus próprios sistemas. As soluções usam diferentes tipos de Sistemas Operacionais, Sistemas de Gerenciamento de Bancos de Dados (SGBD), entre outros componentes de *software* que precisam ser especificados por uma arquitetura computacional adequada para o funcionamento previsto em cada hospital (JENAL; ÉVORA, 2012).

Ao avaliar a evolução da Tecnologia da Informação no InCor - Instituto do Coração da Faculdade de Medicina da Universidade de São Paulo, Gutierrez et al em 2012 destaca a possibilidade da equipe técnica de desenvolvimento de software atuar lado a lado com a equipe assistencial na definição das funcionalidades, alavancando uma visão abrangente dos problemas a serem resolvidos durante a implementação dos sistemas e após, na manutenção do código-fonte e no tratamento dos dados e da informação em Saúde (GUTIERREZ, *[s. d.]*).

Um exemplo de hospital que adota o desenvolvimento do próprio S-RES é o Hospital de Clínicas de Porto Alegre, que desenvolve desde 2014 o sistema AGHUse desenvolvido inteiramente com *software* livre, hoje em uso em hospitais de pequeno, médio e grande porte

e que mantém-se atualizado e aprimorado sustentado pela Comunidade AGHUse. Esta comunidade é composta pelo Exército Brasileiro, a Força Aérea, a UNICAMP, a UFRJ e a Secretaria de Saúde do Estado da Bahia. Nesta comunidade ficam estabelecidos esforços conjuntos para dar manutenção e continuar aperfeiçoando o sistema de gestão hospitalar AGHUse, desenvolvido com tecnologias livres (CHAGAS *et al.*, 2017; DORA *et al.*, 2016; GENRO *et al.*, *[s. d.]*; HCA-FAB, 2018; HCPA, 2014; SANTOS *et al.*, 2016; SILVA, 2019; UNICAMP, 2018; VAZ, 2017).

A EBSERH - Empresa Brasileira de Serviços Hospitalares também optou por instalar em parte de sua rede de hospitais universitários o sistema AGHU, desenvolvido pelo Ministério da Educação entre 2009 e 2014 no Hospital de Clínicas de Porto Alegre a partir da migração do AGH - Aplicativos de Gestão Hospitalar construído originalmente com ferramentas proprietárias e sem atualizações por parte dos fornecedores (ARCENIO, 2015; FLAUSINO; OTHERS, 2015; OLIVEIRA, 2017; RELVA, 2016; SILVA, 2016).

O Ministério da Saúde fornece o e-SUS hospitalar, entretanto esta solução não tem atendido os hospitais, que encontraram dificuldades no seu uso conforme relatado por Silva et. al em 2016 (SILVA; OTHERS, 2016).

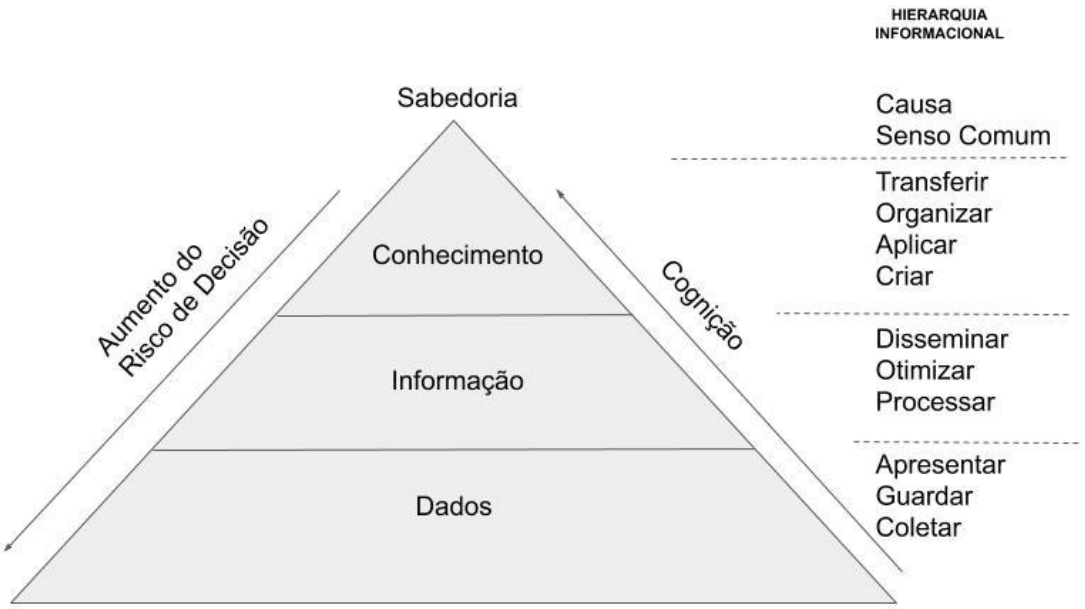
Já o mercado privado de S-RES é composto por dezenas de empresas, que oferecem um grande portfólio de produtos e serviços, com estratégias variadas. Nos EUA o mercado é liderado pelas empresas CERNER e EPIC (RATWANI *et al.*, 2018). No Brasil o mercado de S-RES é liderado pela brasileira MV Sistemas e pela holandesa Tasy-Phillips (BITTAR; BICZYK; SERINOLLI, 2018; SOUZA *et al.*, 2013).

Jenal, S et al em revisão sistemática realizada em 2012 sobre a implantação de sistemas eletrônicos na saúde no Brasil, destaca que a informatização se estabelece em um hospital na medida que o S-RES tem o seu uso efetivo na realização de todo o processamento

dos dados da assistência. Ressaltando que os profissionais da saúde necessitam ferramentas que forneçam informações úteis para a gestão e relevantes para tomada de decisão (JENAL; ÉVORA, 2012).

Em busca desta descoberta de informações, na Ciência da Computação e na Gestão do Conhecimento é muito comum o uso da hierarquia informacional conhecida como Pirâmide DIKW - *Data, Information, Knowledge, Wisdom* (Ver figura 4). Ela oferece um suporte conceitual para a transição que os dados sofrem desde seu estado físico até a formação do que julgamos ser uma informação útil para tomada de decisão e formação da nossa sabedoria (ROWLEY, 2007).

Figura 4 - Esquema gráfico da hierarquia informacional DIKW.



Fonte: Adaptado pelo Autor (2020) do original ROWLEY (2007).

Existe uma sequência lógica que aumenta o valor dos nossos dados na medida em que eles são processados. O dado capturado pode ser enriquecido por tratamento,

preferencialmente, até ele explicar a causa de um fenômeno, ou estabelecer o senso comum de uma explicação. Este tipo de processamento de dados acontece gerando uma coleção de arquivos com grandes volumes de dados que podem ser utilizados em estudos observacionais. (ACKOFF, 1989).

Com o arquivamento destes dados digitais, as grandes bases hospitalares começaram a surgir. A transformação acontece de forma tão profunda que a compreensão das oportunidades em torno dos dados requer uma transição cultural rumo à governança dos dados institucionais. Internalizar as novas oportunidades de melhoria de sistemas demanda educação, engajamento e o mais importante, sucesso visível. Nos Estados Unidos, a Mayo Clinic criou em 2010 um banco de dados chamado "Enterprise Data Trust" para realização de estudos em saúde, e desde então os pesquisadores lá podem consultar dados de diferentes sistemas gerenciados pela Mayo Clinic interconectados através de um modelo para interoperabilidade semântica, permitindo a visualização exploratória das informações de saúde dos pacientes e a descoberta de padrões, amparado pelo uso padronizado de códigos e terminologias da saúde. Funcionalmente, os resultados atualizados em tempo real revelam o impacto de informações bem estruturadas e facilmente consultadas em *queries* sobre eventos clínicos, riscos, resultados e utilização de recursos. Transforma e orienta uma organização de saúde para melhoria da qualidade, produtividade da pesquisa e monitoramento das melhores práticas (CHUTE *et al.*, 2010; WANG *et al.*, 2011).

Outro exemplo de base hospitalar dos Estados Unidos amplamente utilizada para pesquisas em saúde vem do sistema *Veterans Health Information Systems and Technology Architecture* (VistA). Principal ferramenta utilizada em todo o sistema médico do Departamento de Veteranos (VA). O sistema de saúde VA tem mais de 125 hospitais, 800 clínicas ambulatoriais e 135 lares de idosos, todos rodando o VistA. Todas essas instalações

de saúde utilizam o sistema desde 1997 para mais de 8 milhões de veteranos dos EUA. Embora as bases de dados do VA tenha limitações e vieses devido à sua grande porcentagem de pacientes do sexo masculino, o volume de registros de alta fidelidade disponíveis supera essa limitação, tendo em vista que o banco de dados tem sido usado por vários pesquisadores médicos nos últimos 25 anos para conduzir pesquisas de referência em muitas áreas. Os pontos fortes dos dados incluem a capacidade de rastrear todos os medicamentos prescritos que são cobertos pelo sistema VA e a associação desses dados permite que grandes estudos farmacoepidemiológicos sejam feitos com relativa facilidade. (DATA, 2016).

Em 2018 o Prof. Alvin Rajkomar da Faculdade de Medicina da Universidade de San Francisco formou um grupo de pesquisadores para combinar as informações de 2 grandes hospitais das cidades de San Francisco e Chicago. Eles testaram com o apoio da empresa de tecnologia Google novos algoritmos de Inteligência Artificial com grande capacidade de escalabilidade dos serviços, formulando um ambiente com tolerância a falhas para suportar uma grande demanda de requisições com velocidade e alta acurácia nos resultados preditivos, incluindo análises sobre a mortalidade e os diagnósticos médicos (RAJKOMAR *et al.*, 2018).

Entre as bases abertas disponíveis para pesquisa, destaca-se o projeto chamado MIMIC (Medical Information Mart for Intensive Care). Os dados do MIMIC hoje são utilizados por pesquisadores em todo o mundo, promovendo uma ciência aberta e estudos com os dados da saúde em diferentes domínios da ciência. Entretanto, experimentos como o MIMIC também mostram os desafios frente às necessidades para o tratamento dos dados com privacidade e colocam as limitações de análise advindas do processo de anonimização, que precisa ser elaborado sem deixar de ter em perspectiva a utilidade dos dados. Detalharemos o caso do MIMIC no capítulo “Estudos Observacionais no MIMIC”. (JOHNSON *et al.*, 2016b).

A complexidade de um S-RES é proporcional a sua quantidade de requisitos funcionais em uso. Além das funções para assistência, eles também suportam a administração e todos os demais serviços de apoio que um hospital necessita. Os diferentes tipos de S-RES existentes contém um mesmo conjunto de funcionalidades básicas (Ver Tabela 3) . Entretanto, existem diferenciações que incluem módulos adicionais e especializações que variam com a estratégia de negócio ou desenvolvimento de cada S-RES. Cada implementação de S-RES precisa ser adaptada e ajustada em detalhes para os processos assistenciais praticados em cada hospital, permitindo uma grande quantidade de adaptações funcionais e não-funcionais, que podem incluir avanços tecnológicos na infraestrutura de TI, a exemplo da mobilidade, da assinatura digital de documentos e do uso de redes WI-FI para conectar automaticamente o S-RES aos monitores de sinais vitais dos pacientes nas unidades de tratamento intensivo (SILVA, 2016).

Quadro 3 - Módulos comuns em um S-RES.

Módulo	**Informações Tratadas**
Cadastro de Pacientes	Nome, Data de Nascimento, Nome dos Pais, Códigos de Documentos, Escolaridade, Estado Civil, Profissão, Contato (E-Mail, Cel, Tel), Endereços, Códigos dos Pagadores (SUS, Convênio, Particular),
Prontuário Eletrônico	Compêndio de todas as notas médicas, sumários, evoluções e outros registros hospitalares feitos pelas diferentes categorias de profissionais envolvidas na assistência e na administração.
Gestão de Internação	Datas e locais do atendimento do paciente internado. Movimentações do paciente dentro do hospital. Tratamento Intensivo. Isolamento. Censo hospitalar. Painéis Digitais de Informação. Gestão de quartos e leitos.
Gestão de Ambulatório	Datas e locais das consultas e procedimentos ambulatoriais. Clínicas e especialidades médicas. Receitas e encaminhamentos.
Agendamentos	Agendamento e controle de acesso para realizar exames de diagnóstico, consultas, cirurgias, internações, consultas, transplantes, terapias e tratamentos.
Prescrição Médica	Gestão Clínica. Listas de medicamentos, soluções, dietas, diagnósticos, exames, atestados, autorizações e consultorias.
Prescrição de Enfermagem	Gestão do Processo de Enfermagem. Listas de cuidados e diagnósticos de enfermagem.

Registro de Sinais Vitais	Febre, dor, peso, altura, scores de gravidade e prevenção, todos os balanços hídricos e outros sinais e sintomas do paciente, podendo ser automático.
Laudos de Exames	Tipo do exame, nome do exame, item do exame, valores do resultado. Laudo textual do resultado. Datas, locais e motivos da solicitação e para coleta, realização e entrega dos resultados.
Imagens Médicas	Tipo o estudo, imagens do estudo, laudo textual, medições, classificações e diagnósticos.
Farmácia	Lista de materiais e medicamentos, preparo, avaliações e recomendações.
Compras e Estoque	Gestão de almoxarifado. Dispensação de medicamentos. Planejamento de compras. Solicitação de compras. Controle da curva ABC e de qualidade de fornecedores.
Financeiro	Valores financeiros e custos envolvidos na assistência
Administrativo	Cadastros e Relatórios que informam quem, o quê, quando, onde, o custo, como e porquê da produção hospitalar.

Fonte: Elaborado pelos Autores (2022)

3.2. ESTUDOS EM REGISTROS HOSPITALARES SECUNDÁRIOS

Os dados que são coletados pelo próprio pesquisador são considerados Dados Primários de Pesquisa. No caso dos bancos de dados de um S-RES hospitalar, os dados foram coletados com a finalidade assistencial, então dizemos que estes dados são Secundários para Pesquisa. Veremos os diferentes tipos de estudos epidemiológicos que podem fazer o uso destes registros secundários, com diferentes finalidades (HULLEY *et al.*, 2008).

3.2.1 Tipos de Estudos Primários

Existem dois tipos de estudos primários epidemiológicos que podem utilizar bases de registros hospitalares, os estudos experimentais e os estudos observacionais (HULLEY *et al.*, 2008).

Os **estudos experimentais** utilizam consultas (*queries*) realizadas em registros secundários, que podem ser finamente customizadas para filtrar os dados de acordo com critérios pré-estabelecidos, os resultados destas consultas, podem ser utilizados para prospectar e incluir novos participantes em uma pesquisa experimental. Estas consultas, que podem ser feitas de forma exploratória com o apoio de ferramentas de visualização de dados, também podem revelar estatísticas que são importantes para tomada de decisão na definição do delineamento de um estudo experimental, como na definição do número de participantes e na seleção dos métodos de análise apropriados (DATA, 2016).

Entretanto são os **estudos observacionais** que, por definição, se aplicam aos registros hospitalares. Também chamados de pragmáticos, os estudos observacionais servem tanto para avaliar a efetividade e a utilidade de um tratamento em circunstâncias reais observadas durante a assistência, testando hipóteses que podem revelar informações importantes para tomada de decisão, assim como para realizar outros estudos retrospectivos e que podem

revelar fatores de risco para os paciente. Os estudos podem acontecer de forma retrospectiva ou prospectiva. Os principais delineamentos para estudos observacionais são os estudos transversais, os estudo de caso-controle e os estudos de coorte. Suas características e aplicações são apresentadas junto com suas limitações na Tabela (6) do Item 5 deste capítulo. Ainda, os estudos em bases de dados são realizados com baixo custo, se comparado aos Ensaios Clínicos e outros delineamentos experimentais. Permitindo o uso de amostras maiores nas análises e sem demandar o contato com os pacientes (WANG *et al.*, 2016; SOUZA *et al.*, 2015).

3.2.2 Reprodutibilidade e Padronização

Paralelamente ao avanço da informatização na saúde, a comunidade científica está cada vez mais confrontada pela autocrítica feita sobre a falta de reprodutibilidade dos estudos. Com a possibilidade de reproduzir os estudos publicados potencializada pelo compartilhamento de dados e algoritmos, surge o foco editorial de revistas e de jornais científicos em artigos que vêm acompanhados dos dados e do código fonte dos algoritmos necessário para gerar os resultados, tornando o assunto destaque em 2016 na revista Springer Nature (fator de impacto 42.778), em sua sessão Scientific Data ao publicar o artigo de lançamento da base de dados MIMIC e em 2019 no artigo de Gabriel Popkin que utiliza histórias para revelar o poder transformador nos negócios e nas carreiras do compartilhamento de dados entre pesquisadores (COLLINS; TABAK, 2014; JOHNSON *et al.*, 2018; POPKIN, 2019)

Em 2012 a Universidade de Oxford na Inglaterra começou a oferecer uma plataforma de dados para os pesquisadores publicarem seus artigos juntamente com os dados analisados. Assim nasceu o jornal GigaScience de Oxford, (fator de impacto 5.99) e que demanda a

publicação dos dados utilizados em estudos observacionais em sua nuvem especialmente desenhada e construída para o uso em pesquisa (SNEDDON; LI; EDMUNDS, 2012).

Os periódicos do PLOS (fator de impacto 2.7), também utiliza a mesma estratégia de qualificação solicitando que os autores publiquem os dados relacionados aos resultados descritos em seus artigos e as demandas para os autores, invariavelmente solicitam o acesso aos dados analisados sem restrições, no momento da publicação (IENCA *et al.*, 2018).

O jornal Open Health Data apresenta documentos de dados revisados descrevendo conjuntos de dados de saúde com alto potencial de reuso, com repositórios de dados especializados e institucionais, arquivados profissionalmente, preservados e disponíveis abertamente para os cientistas em todo mundo, preferencialmente utilizando padrões de codificação e terminologias (LECUN; BENGIO; HINTON, 2015; OPEN HEALTH DATA, *[s. d.]*; POLLARD *et al.*, 2014; PREVEDELLO *et al.*, 2017)).

Para padronizar as informações na área da saúde, existem especificações de diferentes tipos de modelos e dicionários de dados. O primeiro modelo de dicionário de dados foi definido no IBM *Dictionary of Computing* como um "repositório centralizado de informações sobre dados, contendo metadados que possibilitam dar significado aos dados e documentam os relacionamentos com outros dados, sua origem, finalidade e formato". O estado da arte em dicionários de dados em saúde foi proposto pela OHDSI (pronunciado “Odissey”) que estabeleceu em 2015 uma rede internacional de pesquisadores em bancos de dados de saúde com coordenação da Universidade de Columbia para definir um modelo de dados *open-source* comum para construção de algoritmos que fazem o uso de registros secundários na saúde chamado OMOP. Desde então o OMOP já foi utilizado em dezenas de estudos que podem ser encontrados no site da OHDI (ver Quadro 9). Em estudo para comparar diferentes modelos de dados na saúde, Garza et al. em 2016 na Universidade de Duke apontou as

vantagens do modelo OMOP em relação aos demais, destacando a simplicidade, flexibilidade e aplicabilidade. (GARZA *et al.*, 2016; OHDSI STUDIES, *[s. d.]*)

O modelo OMOP (ver figura 10) foi apontado como estado da arte para pesquisa em saúde por Abrahao MT et al. da USP em artigo publicado pela Sociedade Brasileira de Computação em 2019 destacando a capacidade de customização para o dominio local das instituições no Brasil e também foi utilizado para construir o CDM(Common Data Model) do hospital InCor de Sao Paulo, concluindo que a base de dados final manteve as características dos dados originais extraídos do sistema S-RES do hospital adaptando-se aos padrões do OMOP. (ABRAHÃO; NOBRE; MADRIL, 2019; LIMA *et al.*, 2019)

Em uma revisao de literatura Garza et.al destacou o OMOP entre os diferentes CDMs existentes em experimento realizado na Duke University. Destacou que cada um dos modelos comparados foi desenvolvido para uso específico e sua adequação para outros usos depende de quão próximo o CDM corresponde ao uso planejado de cada hospital. Para esta análise foi definida uma metodologia para avaliação específica de cada CDM com um conjunto de 11 critérios que se enquadram em seis categorias: Cobertura de conteúdo (completude), Integridade, Flexibilidade, Facilidade de consulta (simplicidade), Compatibilidade de padrões (integração) e facilidade e grau de implementação (implementabilidade), tendo o OMO atingido resultados superiores em todas categorias. (GARZA *et al.*, 2016)

As definições e demais informações a respeito dos componentes propostos pela OHDSI podem ser obtidas nas referências apresentadas na lista abaixo.

Lista de recursos para padronização disponibilizados pela OHDSI.

- OMOP Common Data Model, com a definição completa do modelo e as implementações para os diversos bancos suportados https://github.com/OHDSI/CommonDataModel
- Código fonte aberto para instalações: https://github.com/OHDSI/
- Broadsea, ferramenta que disponibiliza uma versão em containers do conjunto de ferramentas, https://github.com/OHDSI/Broadsea
- Repositório Docker com os componentes dockerizados https://hub.docker.com/u/ohdsi/
- OHDSI-In-a-Box que disponibiliza uma máquina virtual pronta para testes e demonstrações: https://github.com/OHDSI/OHDSI-in-a-Box
- Tutoriais e vídeos: Procure no YouTube por OHDSI, existe muita documentação em vídeo (a maioria em inglês).
- Informações gerais: http://www.ohdsi.org
- Fórum de debates: http://forums.ohdsi.org/

Outros modelos existentes a exemplo do Open EHR que propõe a definição de arquétipos e a reutilização de conceitos estabelecidos em "*Templates*" e o padrão de terminologias SNOMED-CT foram propostos em 2011 pelo Ministério da Saúde que publicou a portaria 2073/2011, definindo ambos como padrão para troca de informações entre sistemas S-RES no Brasil. Mas as dificuldades de implementação devido a rigidez dos conceitos desenhados nos agrupamentos de conceitos do domínio no OpenEHR não possibilitaram sua adoção na maioria dos hospitais brasileiros. O OpenEHR não é utilizado nos Estados Unidos e os pesquisadores recentemente vem demonstrando a sua fragilidade em relação ao estado da arte proposto para estudos observacionais e também para a

interoperabilidade em tempo real dos sistemas S-RES dos hospitais que atendem uma mesma população e que propõe o modelo FHIR HL7 como um padrão que pode gerar dados adaptados ao modelo OMOP. Já o SNOMED-CT é amplamente utilizado nos EUA, mas a sua tradução para o portugues do brasil ainda não foi feita e isto afeta a colaboração internacional dos pesquisadores brasileiros, tendo em vista que as bases originadas no Brasil ainda não suportam esta codificação de terminologia. Ainda existem outros modelos internacionais para aplicações específicas na saúde como o STRIDE, proposto para padronização em projetos de pesquisa translacional e todos os demais padrões de codificação existentes somente no Brasil e que usualmente estão em uso nos hospitais por imposição dos sistemas de pagamento do SUS e da ANS (Agência Nacional de Saúde Suplementar), incluindo as terminologias: Terminologia Unificada da Saúde Suplementar (TUSS), Classificação Brasileira Hierarquizada de Procedimentos Médicos (CBHPM) e a Tabela de Procedimentos, Medicamentos, Órteses, Próteses e Materiais Especiais do Sistema Único de Saúde (SIGTAP). (BENSON; GRIEVE, 2016; HRIPCSAK *et al.*, 2015; LOWE *et al.*, 2009; OMOP-CDM CONVERSION AND ANONYMIZATION OF NATIONAL HEALTH INSURANCE SERVICE-NATIONAL SAMPLE COHORT, *[s. d.]*)

3.2.3 Inteligência Artificial em Bases Hospitalares

Com o surgimento destas grandes bases de dados, formadas inicialmente pelo movimento de interoperabilidade dos dados empresariais e governamentais, e posteriormente pela disseminação das redes sociais e dos dados abertos, foram revisitadas as possibilidades de aplicações para Inteligência Artificial nas bases hospitalares com o surgimento do *Machine Learning (aprendizado de máquina)* e do *Deep Learning* (aprendizado de máquina

profundo) (LECUN; BENGIO; HINTON, 2015; PEDREGOSA *et al.*, 2011; POLLARD *et al.*, 2014; SHICKEL *et al.*, 2018).

O aprendizado de máquina com a biblioteca SciKit-learn proposto por Pedregosa et al em 2011 está entre as técnicas mais utilizadas para análise de dados junto ao acervo de bases de dados da OHDSI. Utiliza os paradigmas supervisionado, semi-supervisionado e não supervisionado de aprendizado. Inclui modelos lineares (*Linear Regressions, Bayesian Regression, LASSO*, Perceptron, e outros) e não lineares (*Logistic Regression, Decision Trees, Random Forest, entre outros*) em tarefas regressão e classificação. Destas pesquisas, muitos produtos surgiram para predizer Insuficiência Cardíaca, Hipertensão, Infecções, Re-Admissões, Doenças Mentais e uma série inesgotável de aplicativos para apoiar o diagnóstico e auxiliar na tomada de decisão na saúde. Já o Deep Learning tem como diferencial uma aplicação ortogonal aos paradigmas do *Machine Learning,* quando o aprendizado pode ser supervisionado, não-supervisionado ou por reforço, e pode ser empregado em tarefas, como: *Information Extraction, Representation Learning,* Predição*, Phenotyping* e no que tange a anonimização, também pode ser utilizado para proteção de dados através da desidentificação. Esta mudança de paradigma trouxe uma grande utilidade para inovações que usam os grandes volumes de dados existentes. Em 2019, esta nova tecnologia deu o prêmio Turing para Yann LeCun, Yoshua Bengio e Geoffrey Hinton. O Deep Learning foi apresentado em 2015 na revista Nature e utiliza diferentes redes neurais em módulos parametrizados para treinar com *Backpropagation* uma grande quantidade de dados utilizando *Gradient-Based Optimization.* A area da saude nao demorou para analisar a nova tecnologia e em 2016 o artigo “Deep patient: an unsupervised representation to predict the future of patients from the electronic health records.” foi publicado por Mioto et.al apresentando uma técnica que utiliza autoencoders para predizer desfechos de saúde, com

grande repercussão na revista Nature e tornando-se uma referência na área. (A PRIZE FOR DISCOVERIES PAST, PRESENT AND FUTURE, 2019; LECUN; BENGIO; HINTON, 2015; MIOTTO *et al.*, 2016).

Em 2015 utilizando LSTM Recurrent Neural Networks Lypton et al demonstrou as possibilidades do uso de Redes Neurais na tarefa de diagnosticar pacientes e Choi et al do MIT em 2016 publicou " Doctor AI: predicting clinical events via recurrent neural networks" utilizando um grande volume de dados estruturados e superando significamente a sensibilidade dos métodos anteriores utilizados como *baseline* na predição de diagnósticos, na recomendação de prescrições de medicamentos e na previsão da data e hora da próxima visita. Em 2018 dois grandes hospitais das cidades de San Francisco e Chicago utilizaram novos algoritmos de Inteligência Artificial criados com o uso do Deep Learning pela Google. Utilizando uma infra-estrutura computacional com grande capacidade de escalabilidade dos serviços para suportar uma grande demanda de requisições com velocidade e acurácia nos resultados preditivos sobre a mortalidade e os diagnósticos médicos. Para isso incluíram um total de 216.221 hospitalizações envolvendo 114.003 pacientes clínicos. No momento da admissão, uma admissão média tinha 137.882 tokens (dados discretos), que aumentou durante a permanência do paciente para 216.744 na alta. Para previsões feitas na alta, as informações consideradas em ambos os conjuntos de dados incluídos 46.864.534.945 tokens de dados dos EHRs de ambos hospitais. (CHOI *et al.*, 2015; MIOTTO *et al.*, 2016; RAJKOMAR *et al.*, 2018).

Apesar do *Deep Learning* apresentar desempenho promissor para modelagem preditiva na saúde conforme destacado por Shickel et al da Universidade da Florida em revisão feita sobre aplicações na saúde em 2018, alguns desafios importantes permanecem pois existe insuficiência de dados (ou uma alta frequência de dados missing, o que torna o

tamanho da amostra insuficiente para que os métodos de aprendizado profundo alcancem resultados satisfatórios), e a necessidade de interpretação dos resultados com ferramentas que suportem a de-identificação e a re-identificação dos dados para entregar os benefícios aos pacientes (SHICKEL *et al.*, 2018).

Seguiram os avanços na área da Inteligência Artificial, entre eles o aprendizado de representação em grafos (*Graph Neural Nets*) para recomendação personalizada de *links*, uma tarefa desafiadora não apenas por causa das estruturas complexas de grafos constituídos por vários tipos de nós, elos e atributos, mas também devido à possibilidade de considerar o uso de dados heterogêneos (por exemplo, texto e imagem de uma página na internet) associados a cada nó. As representações estruturadas por métodos de *Deep Learning* devem estar alinhadas com o conhecimento médico e as técnicas que utilizam grafos, como Graph Transformers Networks e Graph Attention Networks são capazes de prover estas explicações eliminando o sentimento de caixa preta das máquinas e permitindo o avanço destas tecnologias na saúde, conforme demonstrados por Choi et al em 2017 que utilizou Graph-Based Attention Model (GRAM), uma técnica que exibe comportamentos intuitivos de atenção ao generalizar adaptativamente para conceitos de nível superior os dados faltantes nos conceitos de nível inferior. Os resultados com o GRAM foram superiores a vários outros métodos preditivos incluindo Recurrent Neural Nets(RNNs) na tarefa de predizer a ocorrência de insuficiência cardíaca. As técnicas para interpretar o resultado de Deep Learning continuam sendo um assunto de relevância para a aplicação na saúde e diferentes métodos têm sido propostos, incluindo: *Convolutional filter response, Output activation maximization, Non-negative matrix factorization* e *Interpretable mimic learning*. Além disto, aspectos qualitativos das bases de dados segue sendo um fator importante para interpretação, informando características sobre dados esparsos, o uso de ontologias e regularizações

urilizadas na formação da base. (CHOI *et al.*, 2017; MA *et al.*, 2018; RAJKOMAR *et al.*, 2018; WANG *et al.*, 2019; YUN *et al.*, 2019; ZHANG *et al.*, 2019)

No campo da Inteligência Artificial, existe uma demanda crescente por algoritmos que tratam as questões de privacidade, frente a demandas práticas. Otimizar o processamento para o tratamento necessário e medir a perda de utilidade da informação ao ser anonimizada, são temas em pleno desenvolvimento científico na área, com diversos estudos explorando o grande manancial de dados disponível nos hospitais ao mesmo tempo em que a privacidade é preservada. (BONAWITZ *et al.*, 2017; MOHASSEL; ZHANG, 2017; XU *et al.*, 2015)

3.2.4 Estudos Observacionais no MIMIC

Para avançar com o uso de base de dados hospitalares com algoritmos de *Machine Learning* e *Deep Learning* possíveis de serem reproduzidos, os pesquisadores do MIT criaram a base aberta MIMIC, com dados de aproximadamente 40 mil pacientes, contendo dados estruturados e textos em inglês para realização de pesquisas tanto nas ciências médicas e da saúde, assim como na estatística, na ciência da computação e na engenharia, sendo referência para mais de uma centena de publicações científicas (ABHYANKAR; DEMNER-FUSHMAN; MCDONALD, 2012; JOHNSON *et al.*, 2016b; POLLARD *et al.*, 2014; SAEED *et al.*, 2002).

Os dados foram coletados no hospital Beth Israel Deaconess Medical Center, localizado na cidade de Boston nos EUA. São diferentes tipos de fontes de dados interconectadas em um banco de dados relacional, que dão origem a uma grande base de registros hospitalares contendo informações dos pacientes que tiveram passagens pela Unidade de Tratamento Intensivo (UTI). Para sua realização, houve um grande envolvimento da academia, da indústria e do hospital, que a cabo conseguiram liberar os dados do MIMIC

em uma plataforma aberta para pesquisadores em todo o mundo (JOHNSON *et al.*, 2016b; MARK, 2016; POLLARD *et al.*, 2014; THE LABORATORY FOR COMPUTATIONAL PHYSIOLOGY, MIT, *[s. d.]*) .

Para dar sustentabilidade ao projeto, foi criada uma comunidade em torno dos dados abertos do MIMIC, que faz o compartilhamento de códigos dos programas de computadores e dos métodos estatísticos utilizados em cada tipo de análise dando suporte aos usuários, incluindo: responder às dúvidas dos usuários conforme necessário, credenciamento de novos usuários, administração da lista de usuários autorizados, criação de conta de usuário, redefinições de senha e concessão e revogação de permissões. Os servidores que fornecem MIMIC incluem autenticação, aplicativos, banco de dados e servidores web que precisam de manutenção periódica. Todos os sistemas devem ser monitorados, mantidos, atualizados e com backup; a carga de manutenção continua a aumentar conforme o número de usuários do banco de dados aumenta. O desenvolvimento e manutenção dos recursos MIMIC foram financiados pelo National Institute of Biomedical Imaging and Bioengineering (NIBIB) e o National Institute of General Medical Sciences (NIGMS) durante o período de 2003 até o momento, unindo esforços da academia (Massachusetts Institute of Technology), da indústria (Philips Medical Systems), e clínica médica (Beth Israel Deaconess Medical Center) com suporte do NIH (National Institutes of Health) . (JOHNSON *et al.*, 2018; MARK, 2016; MIT CRITICAL DATA, 2016).

A base de dados do MIMIC é rica em informações, contendo sinais capturados por sensores multiparamétricos e informações clínicas padronizadas quanto a sua codificação e terminologia. Ao todo são 728.556.685 em 534 colunas (ver Quadro 5), fornecendo 400 bilhões de pontos de dados para o uso em análises de diferentes tipos. (JOHNSON *et al.*, 2018)

Um dos principais objetivos das análises feitas no MIMIC é tentar detectar antecipadamente a degradação do estado de saúde dos pacientes, geralmente acometidos de problemas complexos e graves em tratamento em uma UTI. Além disso, algumas análises realizadas também foram úteis para apoiar na decisão sobre intervenções terapêuticas e para ajudar a melhorar os resultados de outros tratamentos já existentes e no monitoramento dos pacientes.(SAEED *et al.*, 2002)

Quadro 5: Tabela descritiva da base de dados MIMIC.

Tabelas do MIMIC	Colunas	Registros	Comentários
admissions	19	58.976	Internações hospitalares associadas à permanência na UTI.
callout	24	34.499	Registro de quando os pacientes estavam prontos para alta (chamados) e o horário atual da alta (ou mais geralmente, seus resultados).
caregivers	4	7.567	Lista de cuidadores associados à permanência na UTI.
chartevents	15	661.424.966	Eventos que ocorrem em um prontuário do paciente.
cptevents	12	573.146	Eventos registrados na Terminologia Procedimental Atual.
d_cpt	9	134	Dicionário de alto nível da atual terminologia processual.
d_icd_diagnoses	4	14.710	Dicionário da Classificação Internacional de Doenças, 9ª Revisão (Diagnósticos).
d_icd_procedures	4	3.898	Dicionário da Classificação Internacional de Doenças, 9ª Revisão (Procedimentos).
d_items	10	12.487	Dicionário de itens cartográficos não relacionados ao laboratório.
d_labitems	6	753	Dicionário de itens relacionados a laboratório.
datetimeevents	14	4.485.937	Eventos relacionados a uma data e hora.
diagnoses_icd	5	651.047	Diagnósticos relacionados a uma internação hospitalar codificados usando o sistema ICD9.
drgcodes	8	125.557	Estadias hospitalares classificadas usando o sistema de grupos relacionados ao diagnóstico.
icustays	12	61.532	Lista de internações em UTI.
inputevents_cv	22	17.527.935	Eventos relacionados à entrada de fluidos para pacientes cujos dados foram originalmente armazenados no banco de dados CareVue.
inputevents_mv	31	3.618.991	Eventos relacionados à entrada de fluidos para pacientes cujos dados foram originalmente armazenados no banco de dados MetaVision.
labevents	9	27.854.055	Eventos relacionados a testes de laboratório.
microbiologyevents	16	631.726	Eventos relacionados a testes de microbiologia.
noteevents	11	2.083.180	Notas associadas a internações hospitalares.
outputevents	13	4.349.218	As saídas gravadas durante a permanência na UTI.
patients	8	46.520	Pacientes associados à admissão na UTI.
prescriptions	19	4.156.450	Medicamentos prescritos.
procedureevents_mv	25	258.066	Tempos de início e parada do procedimento registrados para pacientes MetaVision.

procedures_icd	5	240.095	Procedimentos relacionados a uma internação hospitalar codificados usando o sistema ICD9.
services	6	73.343	Serviços hospitalares sob os pacientes durante a internação.
transfers	13	261.897	Localização dos pacientes durante a internação hospitalar.
Total	**534**	**728.556.685**	

Fonte: Adaptada pelo autor da original em https://mit-lcp.github.io/mimic-schema-spy/

Uma pesquisa comparou os resultados encontrados do MIMIC frente a outra base de dados aberta, chamada e-ICU, demonstrando em ambas as bases que pacientes de UTI que receberam ventilação invasiva por pelo menos 48 horas estão associados de forma independentemente à maior mortalidade hospitalar (SERPA NETO *et al.*, 2018b).

Também foram realizados diversos estudos de coorte no MIMIC. Uma coorte com 1918 pacientes selecionados avaliou o uso de antibióticos, revelando potenciais efeitos adversos. Em uma outra coorte do MIMIC, a estrutura organizacional das UTIs foi analisada para identificar a influência de algumas decisões sobre a alta do paciente no tratamento de hipertensão nos pacientes, revelando diferenças nos tratamentos recebidos nos dias úteis e finais de semana (WIENS *et al.*, 2018).

Em outro estudo, para avaliar fatores de risco foi identificada associação entre a obesidade e lesões agudas nos rins e que podem levar a morte dos pacientes que estão em tratamento na unidade intensiva. (BOONE *et al.*, 2016)

Foi desenvolvido um pacote para linguagem python chamado TableOne, destinado para padronização da apresentação das estatísticas descritivas do MIMIC e que pode também ser aplicado a qualquer outro programa escrito nesta linguagem. (POLLARD *et al.*, 2018)

Foi criado um visualizador de dados abertos para compreensão de estudos preditivos realizados com Aprendizado de Máquina no MIMIC, com foco na escalabilidade da solução permitindo o uso dos algoritmos por um grande número de usuários. (SHI *et al.*, 2010)

Uma série de eventos de aprendizado, designados como Datathons foram realizados em diferentes universidades utilizando o MIMIC como base de dados para o trabalho. No Brasil, o evento foi realizado no hospital Albert Einstein em 2018. (ABOAB *et al.*, 2016; PIZA *et al.*, 2018; SERPA NETO *et al.*, 2018a)

Inovações em sistemas S-RES foram propostas a partir de testes com os dados no MIMIC para aprimoramento das interfaces que os médicos utilizam para inserir notas em texto livre, propondo uma troca do atual modelo por novas interfaces propostas pela Inteligência Artificial. (SAEED *et al.*, 2002)

A satisfação dos usuários médicos também está no foco dos estudos realizados no MIMIC. Foi proposto um novo algoritmo para organizar as escalas na UTI e assim evitar erros, trocas de plantão e propondo formas de melhorar a satisfação dos médicos, demonstrando resultados promissores e que podem modificar a realidade de um local de trabalho.(RENJIFO, 2005)

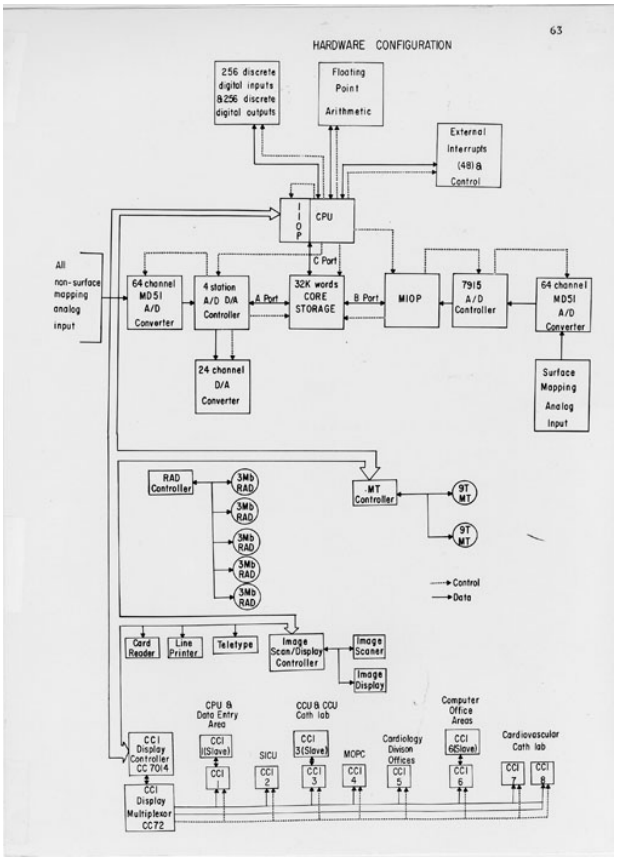
O pesquisador Alistair Jhonson mantém o projeto MIMIC ativo e em 2020 lançou a versão MIMIC-CXR, adicionando novos pacientes na base e incluindo as imagens dos exames de Raio-x realizados durante a internação dos pacientes. Em seu site na internet existe uma grande coleção de artigos e projetos em desenvolvimento em todo o mundo usando o MIMIC. (ALISTAIR'S WEBSITE!, *[s. d.]*; JOHNSON *et al.*, 2019)

3.2.5 Estudos Observacionais no DukeCath

O DukeCath foi um dos primeiros bancos de dados eletrônicos de saúde utilizado em estudos observacionais. Focado em doenças cardiovasculares, o sistema foi fruto da visão do Dr. Eugene Stead, reitor da Duke Department of Medicine entre 1946 e 1967. Uma exibição na internet mostra curiosidades a respeito da sua criação, incluindo os esquemas para

construção da primeira máquina de análise de dados sobre o infarto do miocárdio (Ver Figura 5). (DUKE DATABANK EXHIBIT, *[s. d.]*)

Figura 5. Configuração de hardware de máquina para análise de dados de doenças cardiovasculares.



Fonte: Duke University (DUKE DATABANK EXHIBIT, *[s. d.]*)

Hoje, as bases de registros hospitalares da Duke University nos Estados Unidos são aprovadas para compartilhamento entre os pesquisadores autorizados pela Duke University Health System e pelo Institutional Review Board. Como acontece na maioria das universidades, a solicitação de acesso aos dados é feita de forma individual através de cadastramento, neste caso junto ao Duke Clinical Research Institute. (MENTZ *et al.*, 2014)

Para facilitar este acesso, a universidade criou diferentes conjuntos de dados digitais para os pesquisadores, para o uso em diferentes finalidades, incluindo a realização de pesquisas clínicas e translacionais, assim como fornecendo dados abertos para o treinamento em ferramentas de análise de dados. O conjunto chamado DukeCath contém o registro de 150.000 procedimentos de cateterização em mais de 80.000 pacientes adultos. Os dados foram coletados entre 1985 e 2013, estão desidentificados e são considerados válidos para fins de pesquisa clínica. (FERGUSON *et al.*, 2002)

O objetivo principal dos usuários que realizaram a coleta das informações existentes no DukeCath era fornecer dados para o S-RES em uso na instituição, chamado Epic EHR, utilizado para o atendimento clínico dos pacientes. O registro foi coletado de forma longitudinal, assim cada procedimento de cateterismo está relacionado com dados e características dos pacientes em um determinado momento. Com isto, o seu uso secundário em pesquisas observacionais contém sempre a observação de que os dados do DukeCath incorporam alterações, adições e correções realizadas no registro principal ao longo do tempo. Portanto, pesquisadores que acessarem o DukeCath podem não serem capazes de reproduzir exatamente os resultados das publicações anteriores utilizando dados atuais (KUNTZ *et al.*, 2019; TASNEEM *et al.*, 2017).

Muitos estudos já foram realizados nesta base de dados e sua fundamentação aconteceu ainda na década de 70. Os artigos podem ser acessados no site da Duke Clinical Research Institute. Esta base acompanhou a evolução das técnicas de cateterismo, até a compreensão da influência do índice de massa corporal na eficácia de um tratamento (GRUBB *et al.*, 2020; HARRIS *et al.*, 1979; MENTZ *et al.*, 2014; SHORTLIFFE; PERREAULT, 1990; SOAR DATA™ - DCRI, *[s. d.]*; TURER *et al.*, 2009). Um estudo de

2010 no DukeCath documentou a diferença da chance de sobrevivência em longo prazo entre pacientes de diferentes raças com as mesmas doenças. (THOMAS *et al.*, 2010)

Já o DukeCathR é uma versão deste conjunto de dados para o treinamento e ensino. Os dados foram propositalmente modificados utilizando técnicas estatísticas para anonimizar as informações e assim poder dar acesso amplo a uma base de dados que não pode ser utilizada para pesquisa clínicas, mas podem sim serem utilizados por professores e alunos da universidade em treinamentos sobre análise de dados, desenvolvimento de programas de computador e na realização de experimentos colaborativos. (DUKE, 2016; DUKE DATABANK EXHIBIT, *[s. d.]*)

Os dados disponíveis na DukeCathR incluem o ano em que foi realizado o procedimento de cateterismo cardíaco, demografia do paciente (idade, raça e sexo), histórico (fatores de risco cardiovascular, comorbidades e dias desde procedimento ou evento cardíaco mais recente), sinais vitais e achados do exame físico antes do cateterismo, resultados laboratoriais (creatinina, colesterol LDL, colesterol HDL, colesterol total) com base nos dados mais recentes disponíveis dentro de 1 ano antes do cateterismo e os resultados do cateterismo em si, incluindo a avaliação da estenose nos principais sistemas arteriais coronários, a circulação extracorpórea, enxertos, fração de ejeção do ventrículo esquerdo, se foi realizada intervenção coronária durante o cateterismo e ainda um registro sobre o desfecho do tratamento: morte (todas as causas), infarto do miocárdio não fatal, AVC não fatal, cirurgia de revascularização do miocárdio ou intervenção coronária percutânea. (DUKE, 2016)

3.2.6 Estudos em Bases Governamentais

Entre outras medidas, para formar estas grandes bases de dados que permitem análises epidemiológicas objetivas em registros secundários, mundialmente os governos têm estabelecido mecanismos compulsórios para que os hospitais (e outros estabelecimentos de saúde) façam a notificação de determinados casos para as autoridades, dando origem a diferentes bases de dados com as séries históricas de informação. A qualidade destes dados é avaliada por estudos internacionais, a exemplo do *Global Burden of Diseases* (GBD) *Brazil*, que classificou de 1 a 5 estrelas a qualidade de dados notificados por todos os estados brasileiros. Em todas as regiões do país, ao menos um estado recebeu a nota máxima para qualidade do registro da causa da morte, sendo que toda a região sul ganhou 5 estrelas. (INSTITUTE FOR HEALTH METRICS AND EVALUATION (IHME), 2018; MARINHO *et al.*, 2018)

As bases governamentais em todo mundo são publicizadas com dados processados, com pouco ou nenhum dado individual. Estas bases são fundamentais para os estudos centrados na análise de comorbidades e para a identificação de epidemias e para o controle no desenvolvimento de políticas públicas de saúde, permitindo estudar a distribuição de doenças em diferentes populações. Veremos a seguir uma análise destas bases na Inglaterra e no Brasil.

O *National Health Services* (NHS) é um órgão público inglês, do Departamento de Saúde, uma entidade governamental que tem um dos poucos sistemas de saúde capazes de oferecer um relato completo da saúde em todos os setores de assistência e ao longo da vida para uma população inteira. O Open NHS foi criada em outubro de 2011 com objetivo de fornecer um conjunto de dados de acesso aberto era aumentar a transparência e rastrear os

resultados e a eficiência do setor de saúde britânico. O NHS espera que, ao permitir que os pacientes, médicos e comissários comparem a qualidade e a prestação de cuidados em diferentes regiões do país usando os dados, eles possam identificar de forma mais eficaz e rápida onde a prestação de cuidados está aquém do ideal. É um dos maiores repositórios de dados sobre a saúde no mundo. As informações de alta qualidade capacitem o setor de saúde e assistência social na identificação de prioridades para atender às necessidades das populações locais. Um princípio fundamental é usar o mínimo de dados para satisfazer uma finalidade e eliminar informações relacionadas a um titular de dados que não seja necessário para o processamento específico que está sendo realizado. Este princípio está alinhado com os princípios do NHS e é amparado pelas obrigações de confidencialidade da *Common Law* quanto a Lei dos Direitos Humanos de 1998. Se os registros do paciente forem visualizados de forma identificável, os motivos e o uso dos dados deve ser totalmente documentado e a aprovação é necessária pelo proprietário dos dados apropriado. Isto trilha auditável de acesso aos registros do paciente apóia a Garantia de Registro de Cuidados onde os pacientes devem ser informados sobre quem acessou / viu seus dados e a auditoria fornecer dados precisos em caso de incidentes indesejáveis. Os principais itens a serem documentados são: Quem acessou cada base de dados contendo dados identificáveis; Data e hora de acesso; O motivo do acesso; A saída do acesso. Esta auditoria deve ser mantida em um banco de dados estruturado separado para permitir consultas e auditoria. (OPEN DATA PORTAL (ODP), *[s. d.]*; SHEET, *[s. d.]*)

No Brasil é o DATASUS que mantém estas bases a partir do tratamento dos dados existentes nos 256 sistemas de saúde e de gestão ativos no Ministério da Saúde, incluindo o Boletim de Produção Ambulatorial (BPA), Autorização de Procedimento Alta Complexidade (APAC), Sistema de Informação Ambulatorial (SIA), Sistema de Informação Hospitalar

(SIH), entre outros. (HRIPCSAK *et al.*, 2015; SILVA; DE SOUZA E SILVA; DE AUTRAN, 2019)

Em 2011 o InCor criou um sistema para coletar as diferentes bases de dados do DATASUS para realização de estudos na mineração de dados. A partir da análise de dez anos (período de 2000 a 2009) das bases de dados do SUS. Foram propostos métodos para coleta, limpeza, padronização das estruturas dos bancos de dados, associação de registros aos pacientes do InCor, permitindo a identificação e o seguimento de um paciente com sensibilidade de 99,68% e a especificidade de 97,94% (PIRES, 2011).

Entretanto, em 2014 na Universidade Federal de Alagoas (UFAL) Correia et al. realizou pesquisas em bases de dados bibliográficas, como: PUBMED, Scientific Electronic Library Online (SciELO) e na Literatura Latino-Americana e do Caribe (Lilacs) pelos termos "estudo observacional" em conjunto com "datasus" e foram identificados 972 estudos, sendo ao final revisados 19 estudos realizados especificamente sobre a completude das bases do DATASUS, entre 2005 e 2013. Na revisão foi destacada a restrição de variáveis para acesso público como uma limitação para os estudos observacionais. Nos sistemas SIM e Sinasc é permitido efetuar download dos bancos de dados de todas as variáveis, exceto as de identificação do indivíduo. No Sinan só é possível analisar algumas variáveis pré selecionadas. Os dados da Secretaria de Assistência à Saúde (SAS) e da Secretaria de Vigilância em Saúde (SVS) também fornecem somente parte dos dados coletados, impossibilitando a realização de pesquisas com determinados tipos de escopo. (CORREIA; PADILHA; VASCONCELOS, 2014).

Em outro estudo realizado para avaliar a gestação na adolescência, 15 trabalhos foram selecionados em todo mundo, sendo 6 deles realizados com dados do Sistema de informações sobre nascidos vivos do DATASUS. Os estudos foram realizados entre o período de 2008 a

2012. Dez apresentaram delineamento transversal e 5 estudos de coorte, sendo 4 retrospectivos e 1 prospectivo. Foi identificado que a prevalência de gestantes adolescentes no Brasil foi 26,4% (1.623 casos) bem diferente dos demais estudos internacionais onde a média foi de 10%. (AZEVEDO *et al.*, 2015)

Estudos dessa natureza realizados nas bases de dados do DATASUS podem evidenciar a carência de treinamento dos usuários dos sistemas de informática em saúde que realizam os registros primários durante o processo de assistência ao paciente e a inconsistência de informações que existe em decorrência de atualizações do sistema de informação que não são documentadas juntamente com os dados. Ao aprofundar a compreensão dos resultados destes estudos, fica claro a necessidade de revisão de documentos técnicos referentes à classificação dos tipos de dados e variáveis disponíveis, entre outras informações técnicas importantes para condução de um estudo observacional. (CORREIA; PADILHA; VASCONCELOS, 2014)

Iniciativas que pretendem reorganizar os sistemas de saúde do DATASUS continuam a surgir e a desaparecer. Elas resumem a vontade dos governos em mitigar o resgate de informações incompletas ou inconsistentes nas bases do SUS, e ainda restam definidas em leis e programas de governo, a exemplo do CMD (Conjunto Mínimo e Dados) e do Conecte SUS (PATRÍCIO *et al.*, 2011; PIRES, 2011). De forma mais ampla no Brasil, o Ministério da Saúde entre 2011 e 2019 utilizou uma estratégia chamada e-SUS que visava qualificar os dados e controlar a situação de saúde de forma individualizada da população com a utilização do Cartão Nacional de Saúde (SILVA; DE SOUZA E SILVA; DE AUTRAN, 2019). Entretanto, em Novembro de 2019 o governo federal anunciou uma nova iniciativa relacionada ao uso de dados do governo na saúde, uma estratégia denominada Conecte SUS, que pretende integrar bi-direcionalmente as informações de saúde da população entre os gestores municipais, estaduais e federais e os diferentes estabelecimentos de saúde que

prestam serviços para o SUS, incluindo hospitais, unidades básicas de saúde, unidades de pronto atendimento e outros. “Com o programa, as pessoas vão saber quais vacinas foram aplicadas, os atendimentos realizados, exames, internação", afirmou o ministro em coletiva de imprensa (PROGRAMA QUE VAI INTEGRAR DADOS DE USUÁRIOS DO SUS EM TODO O PAÍS É LANÇADO EM ALAGOAS, 2019).

O sucesso das novas estratégias para melhoria da qualidade dos dados e informações, a exemplo dos experimentos de conexão entre os dados governamentais realizados no CIDACS da FIOCRUZ e que está produzindo uma coorte de 100 milhões de brasileiros, pode estimular o uso integrado destas bases de dados para pesquisa, incluindo no escopo as bases de registro hospitalares que hoje estão isoladas em silos (ICHIHARA; BARRETO; OTHERS, 2017).

3.2.7 Limitações dos Estudos em Bases de Dados Hospitalares

É importante entender as limitações dos resultados apresentados por estudos observacionais em bases de dados hospitalares para interpretar a validade dos achados e assim, atribuir um nível de credibilidade às conclusões de um artigo.

Como características dos estudos observacionais especificamente em bases de dados hospitalares, podemos destacar a complexidade e os custos envolvidos na obtenção, consulta e interpretação correta dos dados de um SGBD de um S-RES. Isto pode demandar o apoio de técnicos especializados em Tecnologia da Informação, Médicos e especialistas de determinados domínio da saúde e ainda, em casos de dúvida, um esforço adicional para validação dos dados eletrônicos contra os documentos originais em papel ou em outros sistemas (ABHYANKAR; DEMNER-FUSHMAN; MCDONALD, 2012; BRAJER *et al.*, 2019; POWELL; BUCHAN, 2005).

Também é preciso considerar o fato de que as bases hospitalares são utilizadas em estudos secundários e não foram desenhadas especificamente para nenhum tipo de estudo, fazendo com que dados importantes possam estar ausentes, inviabilizando determinadas necessidades. (DATA, 2016)

Para utilizar bases hospitalares em estudos epidemiológicos é necessário preservar determinadas características dos dados de acordo com o delineamento do estudo proposto. O tipo de seguimento realizado por cada delineamento interfere no tratamento dos dados (HOCHMAN *et al.*, 2005).

Nos estudos longitudinais precisamos acompanhar o paciente ao longo do tempo e isto implica na preservação da informação do tempo em detalhes para que se possa acompanhar a evolucao da doenca ou do agravo, por exemplo. Nos estudos de coorte e de caso-controle é preciso identificar os pacientes, enquanto indivíduos na linha do tempo, dentro de uma visão longitudinal que permite compreender a etiologia das doenças e também a compreensão dos prognósticos para os tratamento(HRIPCSAK *et al.*, 2015).

Já nos estudos transversais, a comparação é feita antes e depois da exposição, permitindo que as variáveis de data e hora sejam generalizadas para “antes” ou “após”, ou informando o bimestre, trimestre, semestre ou ano do evento de interesse, por exemplo. Dados agregados também podem ser necessários para estes estudos, como: para compreender a morbidade e a mortalidade, ou a efetividade de tratamentos e medicamentos e quais são as reações adversas e interações medicamentosas, agrupando pacientes com determinadas características semelhantes(HOCHMAN *et al.*, 2005) .

Bases de registros secundários podem ser preparadas para fornecer uma ou outra forma de disponibilização dos dados, impondo limitações para determinados objetivos de estudo (Quadro 5). (HULLEY *et al.*, 2008; MIT CRITICAL DATA, 2016)

Quadro 5: Delineamentos de pesquisa e exemplos de aplicações de bases de registros hospitalares em estudos epidemiológicos (elaborada pelo autor)

	Delineamento de pesquisa	**Tipo de estudos em bases hospitalares**	**Limitações**
1	Estudos observacionais de casos e relatos de casos	Quadro clínico dos pacientes e estudos de doenças raras	Estudos descritivos, são limitados a leitura e interpretação de todo o prontuário dos pacientes, envolvendo consulta e acesso a dados de todos os tipos e atributos, tornando a captura dos dados complexa e demorada para obtenção de um pequeno número de pacientes .
2	Estudos observacionais analiticos transversais	Estimar a frequência de um determinado evento. Analisar a associação entre exposição e desfecho. Comparar a prevalência do desfecho entre expostos e não expostos Razão de prevalência	Característicos por não haver temporalidade (importante para estabelecer causalidade). A exposição e o desfecho são coletados da base de dados sem a informação de datas associadas aos eventos analisados. Não se pode afirmar o que foi causa ou consequência (causalidade reversa).
3	Estudos observacionais analiticos de caso-controle	Identificar fatores de risco em doenças raras ou de longo período de latência. Analisar epidemias Estudar a etiologia das doenças. Razão de chances.	Necessitam de longos períodos de coleta de dados de forma sistematizada e os estudos podem demandar conjuntos de dados com anos de coleta, incluindo confundimentos derivados das mudanças na forma e uso dos S-RES ao longo dos anos..
4	Estudos observacionais analiticos de coorte	Avaliar os riscos e benefícios do uso de determinada intervenção. Estudar a incidência da doença entre os expostos e não expostos às mudanças. Comparar a incidência usando uma proporção (risco relativo) ou uma diferença (risco atribuível). Estudar a evolução e o prognóstico das doenças.	Estudos de coorte retrospectivos: limitados aos dados de saúde observados na base hospitalar, sem registrar a integralidade dos eventos de saúde do paciente. Estudos prospectivos: Somadas às limitações dos estudos retrospectivos, também estão limitados a necessidade de interoperabilidade entre a base de dados tratada pelo pesquisador e o S-RES dos hospitais, demandando a automatização do preparo dos dados.
5	Estudos Experimentais	Seleção de participantes através da pesquisa por critérios na base de dados para estudos de novos tratamentos e formas de prevenção.	Critérios ambíguos nas bases de dados podem incluir novos tipos de vieses de seleção nos estudos experimentais.

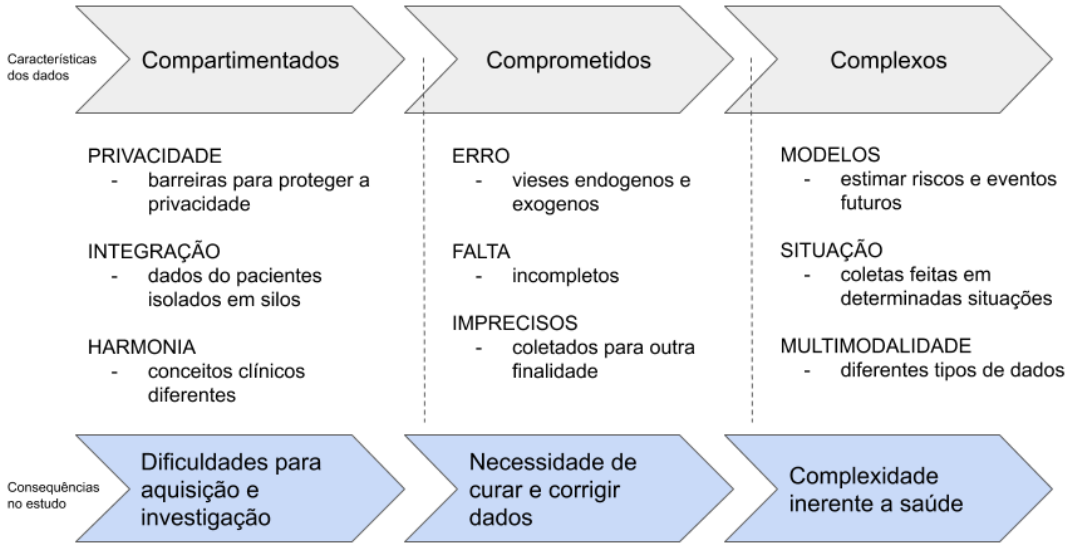
Fonte: Elaborado pelo autor (2020).

Outro ponto destacado pelos pesquisadores do DukeCath é de que os estudos retrospectivos em registros secundários estão sujeitos a mudanças na qualidade e na forma como os dados são coletados, podendo demandar a necessidade de traduzir os dados de uma

época para outra, e também inviabilizando a reprodução de resultados obtidos no passado ao utilizar os dados mais atuais (DUKE, 2016).

Os dados da saúde são fragmentados e por isto, os estudos de Coorte mesmo sendo potencializados pelo uso de grandes bases hospitalares encontram limitações no seu uso. Pessoas saudáveis não fazem testes e exames de diagnóstico com frequência e têm pouco ou nenhuma informação de saúde registrada longitudinalmente. A atenção primária tem parte dos dados com as informações de saúde dos pacientes, incluindo vacinas e consultas médicas e estas raramente estão presentes nos S-RES hospitalares. Médicos especialistas têm em suas clínicas dados de evolução de doenças específicas (cardiologista, dermatologia, etc). As farmácias têm os dados da frequência de compra e uso dos medicamentos pelos pacientes. Diferentes hospitais têm fragmentos de dados multidisciplinares sobre a evolução das doenças e dos tratamentos realizados pelos pacientes. Laboratórios têm dados de exames, que são feitos em pacientes saudáveis e doentes. Nenhuma fonte de dados tem todos os dados de um paciente e além disto, elas estão expostas a uma grande quantidade de vieses conforme destacado pelo pesquisador do MIT Alistair Johnson em artigo dedicado ao estudo destas limitações inerentes ao uso de bases de dados hospitalares (Figura 8). (MIT CRITICAL DATA, 2016; REPS *et al.*, 2018)

Figura 8. Limitações e desafios para os pesquisadores.



Fonte: Adaptado pelo autor do original em (JOHNSON *et al.*, 2016a).

As informações existentes em registros secundários estão expostas a todos os tipos de vieses, sujeitas a baixa qualidade dos processos, contendo erros de digitação ou medições feitas de forma errada e por isto exigem tratamento apropriado para a sua análise. Estas limitações precisam ser reconhecidas e percebidas pelos pesquisadores que utilizam estes dados. (KHOURY; IOANNIDIS, 2014).

Apesar dos avanços, as limitações destes estudos têm sido apresentados em diferentes revisões sistemáticas sobre estudos centrados na análise de bases de dados elaborados a partir de Registros Eletrônicos de Saúde (POWELL; BUCHAN, 2005).

Em 2013, Freeman et al realizaram uma revisão sistemática da literatura publicada sobre o uso de base de registros secundário para vigilância de infecções hospitalares. A implementação dos sistemas de vigilância eletrônica foi viável em muitos cenários distintos e

os resultados sugerem que os sistemas hospitalares já produzem dados com qualidade e quantidade suficientes para treinar modelos de dados que aprimoram a eficácia do monitoramento dos pacientes, mas destacam que nenhum dos estudos teve sua validade externa analisada, limitando o potencial dos modelos criados (FREEMAN *et al.*, 2013).

No escopo da LGPD, no que tange a anonimização dos dados para a pesquisa acadêmica com dados observacionais, existe uma indicação de que a prática de anonimizar os dados deve ser exigida sempre que possível, deixando em aberto os requisitos específicos e técnicos para regulamentação futura e que podem trazer novas limitações a este tipo de estudo (BRASIL, 2018).

O Ministério da Saúde especificou o Conjunto Mínimo de Dados da Atenção à Saúde (CMD), que juntamente com a documentação oficial do Cadastro Nacional de Estabelecimentos de Saúde (CNES), o Repositório de Terminologias em Saúde (RTS) e o Sistema de Gerenciamento da Tabela de Procedimentos, Medicamentos, Órteses, Próteses e Materiais Especiais do SUS (SIGTAP) definem as especificações para codificação e padronização dos dados da saúde no Brasil e os problemas existentes nestas bases também podem impor limitações aos estudos. (PORTAL CMD 1.0, *[s. d.]*)

O CMD, em especial, vai descrever procedimentos para coletar os dados de todos os estabelecimentos de saúde do país que usam um S-RES e de forma integrando ao Sistema Nacional de Informação de Saúde (SNIS). A implantação do CMD está acontecendo de forma modular e gradual no Brasil e ainda não existem planos para o seu uso integral em todo país. (Ver Figura 8). Em uma revisão sistemática de 2019 evidencia que ainda não existem sistemas S-RES utilizando todas as terminologias definidas pelo DATASUS impondo limitações severas a condução de estudos epidemiológicos no Brasil que integram dados de diferentes origens (MACIEL; FERREIRA; DE FÁTIMA MARIN, 2019).

3.4. ANONIMIZAÇÃO

A anonimização de dados é definida como um processo capaz de alterar os dados de forma irreversível objetivando a privacidade. De forma determinística, trata da certeza da proteção dos dados pessoais, prevendo as diferentes formas de reidentificação possíveis em microdados (dados dos indivíduos) e dados tabulares (agregados) (WILLENBORG; DE WAAL, 2012).

Em relação às leis de proteção de dados em todo mundo, incluindo a LGPD no Brasil, uma vez anonimizados os dados não são mais passíveis de aplicação da lei e assim podem ser compartilhados de forma livre entre diferentes partes. Entretanto, para o uso de dados na pesquisa a anonimização deve ser utilizada sempre que possível e com as técnicas razoáveis para aplicação. Dados anonimizados que preservam uma chave específica que possibilita a re-identificação, se necessário, são considerados pseudo-anônimos e não são enquadrados pelas legislações como dados anônimos. Para garantir o anonimato o tratamento precisa ser irreversível (BRASIL, 2018).

No mercado empresarial, os dados anônimos usualmente são utilizados para elaboração de análises de negócio que envolvem diferentes departamentos e instituições, motivados pela possibilidade de mitigar riscos relacionados a quebra de privacidade com seus clientes, governos e outros interessados no compartilhamento de informações (GREENBERG *et al.*, 2016).

Na Estatística, diferentes técnicas de proteção de dados (Ver Quadro 12) são tratadas na disciplina de SDC - Statistical Disclosure Control (Controle de Privacidade Estatístico), quando a anonimização deixa de ser uma característica dos dados e passa a ser tratada como uma propriedade estatística que pode ser medida e assim, a perda da informação no

processamento necessários para anonimizar os dados pode ser compreendida. (WILLENBORG; DE WAAL, 2012)

O tratamento para anonimização por algoritmos é complexo e usualmente cada algoritmo é criado utilizando uma combinação de métodos de anonimização existentes. Cada método a priori precisa da definição dos tipos de dados envolvidos (vistos no Capítulo “Tipos de Dados”). Também demanda a definição dos atributos de privacidade (quais são os dados sensíveis no conjunto), um modelo de privacidade que pode ser preparado para suportar diferentes tipos de ataques de adversários, incluindo: ataques de “data linkage” onde adversários que são detentores dos dados (por exemplo: quando um membro do projeto de pesquisa tenta identificar um paciente) tentam ligar bases de dados através de seus quasi-identificadores. Fica ainda mais complexo tratar a anonimização, quando os dados são expostos a adversários atuando de forma conjunta com qualquer outra pessoa ou instituição (no âmbito digital dos dados abertos na internet e no âmbito real dentro ou fora das empresas) ou de adversários que conhecem a distribuição dos dados anonimizados a priori e podem até mesmo possuir conhecimento prévio a respeito de valores sensíveis existentes em determinados registros. Por fim, todos os métodos de proteção de dados precisam estabelecer métricas de informação para aferir a utilidade dos dados após o processamento.

Para estabelecer um domínio específico para tratarmos o assunto anonimização na área da saúde, vamos propor uma adaptação e uma extensão para a área da saúde no Brasil da ontologia proposta por Queiroz et al no XII Brazilian Symposium on Information Systems, evento da SBC - Sociedade Brasileira de Computação que aconteceu no ano 2016 em Santa Catarina, definindo classes e subclasses de informações sobre a preservação da privacidade em dados publicados pelo governo brasileiro. A adaptação proposta define uma nova classe para os tipos de estudos epidemiológicos dentro da ontologia e a extensão inclui dentro dos

modelos de privacidade propostos, outros métodos encontrados na literatura além da supressão e da generalização, a exemplo da privacidade diferencial e da sintetização de dados. (QUEIROZ; LINO; GUSTAVO H M, 2016).

3.4.1 Atributos de Sensibilidade

No âmbito da anonimização, os dados podem ser classificados em um dos seguintes atributos de sensibilidade (FUNG *et al.*, 2010; SWEENEY, 2002a).

- **Identificadores**: dados explícitos de identificação de uma pessoa natural (nome completo, CPF, registro profissional)
- **Quasi-Identificadores**: quando combinados podem revelar a identidade de dados pessoais (local e data de nascimento em conjunto com sexo e etnia).
- **Dados Sensíveis**: informação privada e/ou confidencial (doença)
- **Dados Não-Sensíveis**: informações que não colocam em risco a privacidade do paciente caso sejam identificadas e não são classificadas em nenhum dos demais atributos anteriores.

Outros autores utilizam uma nomenclatura diferente para definição destes atributos, como é o caso de Martinez et al que em sua tese utiliza *Identifier, Confidential e Non-Confidential* para se referir a *Explicit Identifier, Sensitive Attribute* e *Non-Sensitive Attribute*, respectivamente (FUNG *et al.*, 2010; MARTÍNEZ LLUÍS; OTHERS, *[s. d.]*).

3.4.2 Métodos de proteção de dados

A reidentificação é uma ciência relacionada ao estudo das técnicas para identificar dados considerados anônimos pelo tratamento com diferentes técnicas de proteção de dados

(Ver Tabela 6). Entretanto, a exposição digital da privacidade nas redes sociais e nos portais colaborativos, somadas a abertura dos dados governamentais e o surgimento de bases abertas não governamentais, vem trazendo um grande número de desafios para os pesquisadores desta área. Segundo a pesquisadora L. Sweeney, a grande maioria dos dados considerados hoje anônimos podem ser identificados utilizando ou a abordagem jornalística de investigação, que conecta peças externas a um conjunto de dados específico, ou utilizando a abordagem do cientista de dados, que utiliza modelos estatísticos para evidenciar itens que possam interconectar diferentes bases de dados e assim retroceder o processo de anonimização. (ROCHER; HENDRICKX; DE MONTJOYE, 2019; SWEENEY, 1997).

Quadro 6 : Definições de diferentes métodos para proteção de dados.

Método	**Descrição da proteção de dados**	**Fonte:**
Anonimização	Remoção irreversível do vínculo entre o indivíduo e seus dados na medida em que seria impossível restabelecer uma ligação.	(BRASIL, 2018; KUSHIDA *et al.*, 2012; QUEIROZ; LINO; GUSTAVO H M, 2016; SWEENEY, 2002b)
Desidentificação	Remoção ou substituição de identificadores pessoais para tornar difícil restabelecer um vínculo entre o indivíduo e seus dados; regra de privacidade da HIPAA.	(SULLIVAN, 2004; SZARVAS; FARKAS; BUSA-FEKETE, 2007)
Pseudo Anonimização	Os dados de identificação na pseudo anonimização são transformados e substituídos por um especificador que não pode ser associado aos dados sem conhecer uma chave de identificação.	(BRASIL, 2018)
Despersonalização	Processo de identificação e separação dos dados pessoais de outros dados	(KUSHIDA *et al.*, 2012)
Criptografia	Processo para tornar os dados secretos (exceto para o destinatário pretendido) usando um algoritmo.	(BRACCI; CORRADI; FOSCHINI, 2012)
Função hash	Algoritmo que converte um grande conjunto de dados em um pequeno dado, geralmente um único número inteiro que pode servir como um índice alternativo.	(BRACCI; CORRADI; FOSCHINI, 2012)
Ofuscação	Mudança do significado pretendido no registro da informação, tornando a interpretação confusa, intencionalmente ambíguo e mais difícil de interpretar.	(BRACCI; CORRADI; FOSCHINI, 2012)
Supressão	Remover dados que possam identificar diretamente uma pessoa.	(KUSHIDA *et al.*, 2012)
Generalização	A generalização representa objetos do mundo real que possuem os mesmos atributos e que podem ser categorizados. Pode definrir uma hierarquia que mostra as dependências entre entidades de uma mesma categoria.	(SWEENEY, 2002b)
Perturbação	Inclusão de ruído ou sujeira de forma intencional nos dados para dificultar a re-identificação.	(DWORK; ROTH, 2014)

Privacidade Diferencial	Técnica de anonimização de grandes volumes de dados que utiliza a perturbação de dados para gerar dados com atributos de segurança gerenciados.	(DWORK; ROTH, 2014)

Fonte: elaborado pelo autor (2022)

3.4.3 Modelos de Privacidade

A pesquisadora L. Sweeney afirmou em 1997 que as técnicas existentes para remoção de PHIs dos conjuntos de dados nos EUA são insuficientes para garantir a anonimização de dados, em especial nos dados da saúde. Afirmando que todos os pacientes presentes em bases de dados desidentificadas (mas não anonimizadas) estariam a mercê de serem re-identificados caso estas bases forem liberadas na internet. Estes pacientes poderiam até mesmo serem contatados, sendo 87% da população americana unicamente identificável com o uso de apenas 3 variáveis (data de nascimento, sexo e CEP) (SWEENEY, 2002b, 2015).

Entre os diferentes métodos de proteção, a supressão de dados é a maneira mais rápida de remover identificadores diretos, como nome e número do passaporte ou por quasi-identificadores, como: sexo, cep e data de nascimento. Foi demonstrado que 66% dos norte-americanos podem ser reidentificados somente com estas três variáveis. A supressão também pode acontecer com a eliminação de valores outliers, ou de qualquer outro valor que possa identificar um conjunto restrito de pessoas.(GOLLE, 2006).

Os principais métodos determinísticos de proteção de dados para atingir a anonimização, compõem a família de algoritmos k-anonymity, l-diversity, t-closeness e suas extensões.

- k-Anonimity: o método parte da uma premissa na generalização e supressão de quase-identificadores que garantem que os dados de um indivíduos são indistinguíveis de k-1 outros indivíduos. (SWEENEY, 2002b).
- l-diversity: o custo computacional reduzido em relação ao k-anonymity em função da generalização dos dados sensíveis, que aumenta a complexidade na medida em que uma hierarquia precisa ser definida para cada valor sensivel.
- t-closeness: Protege contra os ataques de conhecimento prévio, uma fragilidade conhecida do k-anonymity e contra a identificação da proximidade semântica dos atributos, uma limitação do l-diversity.
- Outros: Extensões aos métodos da família podem fazer a realocação, que acontece quando é necessário perturbar os quasi-identificadores dos outliers modificando os valores reais. O modelo δ-Presence pode ser usado para proteger os dados da divulgação de membros (membership disclosure), onde um conjunto de dados revela a probabilidade de um indivíduo da população estar contido no conjunto de dados. O método β-Likeness visa superar as limitações dos modelos anteriores, restringindo a distância máxima relativa entre as distribuições de valores de atributos sensíveis, considerando também o ganho de informação positivo e negativo.

O algoritmo **k-anonymit**y proposto por Sweeney et al. usa generalização e supressão para evitar estes ataques. Em geral, um valor original é alterado por outro que apresenta a informação agrupada, mas sintaticamente o valor permanece constante. Quando a anonimização é alcançada somente com o uso da generalização, em casos de multi-dimensionalidade corre-se o risco de super generalizar a informação, comprometendo

sua utilidade. Também faz o uso da supressão, quando os valores são representados como um asterisco '*'. Assim o poder de vincular uma pessoa e distinguir dados através de quasi-identificadores fica restrito quando é utilizado o k-anonymity. Alguns pontos positivos deste algoritmo podem ser destacados pois ele preserva a divulgação de identidade inibindo a conexão (linkage) para um conjunto de dados com valores menores que 'k' indivíduos, impedindo que o adversário conecte um elemento sensível a dados externos. O custo computacional incorrido no estabelecimento desse método é consideravelmente menor em comparação com o custo de outros métodos de anonimização. Existem diferentes implementações abertas destes algoritmos, como Datafly e o Mondrian, que são usados extensivamente. Entretanto, este algoritmo apresenta limitações conhecidas e que demandam adaptações ao método para proteger ataques do tipo “correspondência não classificada”, quando a heterogeneidade nos atributos sensíveis é inadequada e gera pequenos agrupamentos que expõem informações, ou quando um adversário tem um conhecimento prévio sobre o indivíduo e com raciocínio lógico adicional ou apoio de dados externos identifica seus atributos sensíveis e faz a re-identificação. (GIONIS; TASSA, 2009; RAJENDRAN; JAYABALAN; RANA, 2017; SWEENEY, 2002b).

Para contornar essas limitações, foi criado o algoritmo **l-diversity**. Funciona como uma uma extensão do antecessor e introduz um novo método que pode garantir a privacidade dos dados mesmo quando o adversário tiver conhecimento prévio sobre atributos sensíveis. Esta abordagem funciona em torno da percepção de como os atributos sensíveis em cada grupo estão representados, geralmente com o uso de generalização. Outro ponto a favor no l-diversity em relação ao k-anonymity é o custo computacional reduzido em função da generalização dos dados sensíveis. O principal ponto negativo deste algoritmo é a sua complexidade, que demanda trabalho adicional para definição de hierarquias para os dados

sensíveis, e sua fragilidade contra ataques de assimetria e ataques de similaridade, quando é possível atribuir uma exposição devido ao relacionamento semântico entre os atributos sensíveis generalizados. (RAJENDRAN; JAYABALAN; RANA, 2017)

O algoritmo **t-closeness** foi proposto como uma melhoria do l-diversity, propondo uma técnica que calcula a distância entre os valores e diminui a granularidade dos dados interpretados. Ele reduz a correlação entre os atributos quase-identificadores e os atributos sensíveis propondo uma extensão do conhecimento do adversário, definindo que seu conhecimento sobre os dados específicos é limitado, mesmo quando este conhecimento não é restrito ao conjunto dos dados anonimizados. Protege contra os ataques de conhecimento prévio, uma fragilidade conhecida do k-anonymity e contra a identificação da proximidade semântica dos atributos, uma limitação do l-diversity. (PRASSER; KOHLMAYER, 2015)

3.4.4 Privacidade Diferencial

Atingir a anonimização de determinados conjuntos de dados é um processo desafiador, ainda mais no âmbito conectado pela internet. A técnica mais avançada para promover a privacidade e a proteção de dados digitais em situações que podem revelar hábitos de uso e consumo de produtos digitais se chama Privacidade Diferencial.

Estudos relacionados a re-identificação vêm chamando a atenção da comunidade científica acerca das questões de privacidade e revelando a importância da privacidade diferencial. Cinthya Dwork propôs em 2006 o conceito E-Differential Privacy, uma nova maneira para obter dados que preservam a privacidade, definindo uma função estocástica do mecanismo, ao invés de um resultado pragmático. Com isto, qualquer conjunto de dados processado pelo mecanismo pode ter o mesmo nível de privacidade dos demais conjunto de dados, não importa quão extremo ou sensível o conjunto de dados seja. A premissa por trás

da privacidade diferencial é de que uma pessoa não será afetada adversamente ao ter seus dados utilizados em qualquer tipo de estudo, independentemente de outros conjuntos de dados ou fontes de informação que possam existir. (ABADI *et al.*, 2016; DWORK, 2006)

Diferente dos demais mecanismos de proteção de dados, ela não é um processo irreversível mas sim um método estocástico que trata das chances da reidentificação. É uma propriedade dos conjuntos de dados gerados por algoritmos que permitem identificar, medir e definir a probabilidade que os dados têm de serem realmente anônimos e protegidos quanto a qualquer tipo de ataque para desanonimização e quebra de privacidade. (DWORK, 2006)

Para a obtenção da privacidade diferencial, a anonimização é alcançada através da inclusão de ruído através da perturbação de valores. O resultado são dados úteis para finalidades muito específicas, como por exemplo: realizar enquetes privadas sobre posicionamento político nas redes sociais, ou coletar informações do usuário de modo silencioso para aprimoramento dos computadores da Apple e do navegador de internet Google Chrome. Demonstrado por Dwork et al em um estudo que abordou brechas de privacidade no serviço de administração de banco de dados de uma grande empresa, que ao fornecer mais centenas de vezes a mesma Query SQL reprocessada com novos dados, revelou que a probabilidade posteriori do valor de um atributo em um registro ser igual a 1 era diferente da sua probabilidade a priori, podendo assim revelar o risco de um indivíduo fazer parte ou não do conjunto de dados, por exemplo, das promoções salariais de uma empresa. (DWORK; ROTH, 2014; EL EMAM; DANKAR, 2008)

Tanto a relocação quanto a perturbação utilizada na privacidade diferencial envolvem a modificação dos dados e isto traz um grande prejuízo, não só relacionado a perda de utilidade, mas para percepção de confiança na informação. (EL EMAM; DANKAR, 2008)

3.4.5 Sintetização de Dados

Uma alternativa para a anonimização é a sintetização, que produz dados realísticos, mas que não são reais. Estudos recentes apontam para a inviabilidade de estudos clínicos nestas bases de dados, ao mesmo tempo em que elas são apontadas para o uso em treinamento de ferramentas e técnicas de análise (DAHMEN; COOK, 2019; HOLOHAN *et al.*, 2017; RAJENDRAN; JAYABALAN; RANA, 2017)

Uma técnica utilizada para gerar dados sintéticos com o uso de Inteligência Artificial é a MICE - *Multiple Imputation Chained Equations*, que implementa uma cadeia de algoritmos de *Machine Learning* classificadores treinados com aprendizado supervisionado de dados reais, para imputar dados sequencialmente e progressivamente adicionando novas colunas de acordo com o dicionário de dados da base original e que dará origem a novas bases de dados sintéticos. (AZUR *et al.*, 2011; GONCALVES *et al.*, 2020)

A perda de confiança é o grande problema por trás da técnica de sintetização de dados, que utiliza algoritmos para produzir dados de pacientes hipotéticos a partir dos dados de pacientes originais, formando grandes conjuntos de dados mas que não tem utilidade para pesquisa clínica. (CHEN *et al.*, 2019).

3.4.6 Desidentificação de Texto Livre

Em 2007 Szarvas et al publicou um estudo afirmando que “o anonimato dos registros médicos é de grande importância, porque um texto não identificado também pode ser disponibilizado ao público para facilitar a pesquisa sobre doenças humanas”. A necessidade de compartilhar dados e reproduzir experimentos está no âmago da ciência, revelando a

importância dos métodos algoritmos de tratamento de dados para anonimização (SZARVAS; FARKAS; BUSA-FEKETE, 2007)

Para variáveis não estruturadas do tipo texto livre, as técnicas básicas de anonimização envolvem o uso de Processamento de Linguagem Natural (PLN) para construção de algoritmos que funcionam através da formulação de dicionários que podem ser compostos por listas obtidas através de fontes abertas na internet ou por catálogos da própria instituição onde os dados tiveram origem. Estes dicionários podem conter os nomes das pessoas envolvidas, assim como listas de números de licenças profissionais, localizações geográficas, nomes de ruas das cidades onde residem pacientes e funcionários da instituição, cidades da região, estados e países, o nome das maiores cidades do mundo, nomes de doenças e uma lista contendo todas as informações que são consideradas Informações Pessoais de Saúde, ou Personal Health Information (PHI) conforme definição na legislação dos EUA (SNELL, 2017).

Os primeiros algoritmos considerados o estado da arte na anonimização utilizavam métodos que usavam recursos ortográficos. Os algoritmos contavam letras maiúsculas, tamanho da palavra, informações comuns sobre a forma da palavra (contém um dígito ou não, possui caracteres maiúsculos dentro da palavra, possui sinais de pontuação na palavra, possui dígitos na palavra, o número é romano ou não) e várias expressões regulares que descrevem as características comuns dos campos de identificação, como por exemplo o formato dos números de telefone celular com dois dígitos entre parênteses para identificar o código de área local. Também eram utilizadas informações sobre a frequência das palavras de um *corpus* composto por textos coletados de outras fontes, assim como a frequência de uma palavra dentro de um único documento. Outras técnicas calculavam a proporção das ocorrências entre maiúsculas e minúsculas, a proporção de frequências em maiúsculas no

início de uma frase e a detecção de informações frasais: como uma classe prevista por várias outras palavras anteriores e a presença de sufixos de frases comuns. Informações contextuais também foram agregadas as técnicas de anonimização baseada em recursos ortográficos, incluindo a posição da sentença no texto, o cabeçalho da seção mais próxima, palavras do texto que geralmente precedem ou seguem um PHI, se a palavra caiu entre aspas, se a palavra caiu entre colchetes, ou ainda se todo o contexto está em maiúsculas. (SZARVAS; FARKAS; BUSA-FEKETE, 2007).

No MIT foi desenvolvido o algoritmo chamado "deid de-identification package" para realizar o trabalho de retirar os dados de identificação dos pacientes das bases de dados do laboratório physionet. Em 2004 o algoritmo foi utilizado para elaboração do MIMIC II e foi licenciado abertamente junto com o seu código fonte, disponível livremente. Ele foi desenvolvido e testado usando um corpus de texto livre como padrão ouro, contendo 2.434 notas de enfermagem que foram completamente desidentificadas por um processo com várias etapas de tratamento, incluindo uma variedade de métodos automatizados, que exigiu revisões detalhadas feitas por três especialistas trabalhando de forma independente. Quando um PHI é encontrado pelo deid nas anotações de enfermagem, eles são substituídos por marcadores que podem identificar o tipo do PHI encontrado, ou por dados substitutos realistas e que aumentam a utilidade da base desidentificada. Embora o software deid possa ser redistribuído nos termos da *General Public License* (GPL), o seu corpus padrão ouro, devido à possibilidade muito pequena de conter uma ou mais instâncias de PHI ainda não detectada, está disponível atualmente apenas para aqueles pesquisadores que possuem acesso autorizado aos bancos de dados clínicos PhysioNet. (DE-IDENTIFICATION SOFTWARE PACKAGE V1.1, 2007)

A competição de análise de textos da saúde com textos anonimizados realizada pelo centro nacional de pesquisa em bioinformática I2B2 - Informatics for Integrating Biology and the Bedside ajudou a aprimorar o estado da arte na desidentificação de texto livre. Promovida anualmente, a série de desafios mantida pela Harvard Medical School foi recentemente renomeada para N2C2 e agora fornece o acesso a dados de forma controlada por meio de um portal especializado. Os resultados do N2C2 são apresentados em um Simpósio Anual e posteriormente publicados. (BOUSSADI; ZAPLETAL, 2017; DE BRUIJN *et al.*, 2011; LUO *et al.*, 2018)

3.4.7 Utilidade e Perda de Informação

Para orientar a coleta, identificar padrões, testar hipótese e interpretar resultados, os cientistas estabeleceram na Estatística os seus esforços para desenvolver o pensamento descritivo e inferencial. Assim, funcionando como uma analogia a "luz da ciência moderna", é a estatística que tem servido de suporte para tomada de decisão do pesquisador em relação a sua hipótese de pesquisa. Com isto o compartilhamento e a disseminação eficazes de registros hospitalares pode ocorrer apenas se o detentor dos dados tiver garantias de que a divulgação de informações confidenciais não representa um risco a privacidade, ou um risco para a qualidade das decisões que serão suportadas pela análise dos dados (CIRIANI *et al.*, 2007).

A perda de informação é a quantificação da distorção imposta aos dados como parte do tratamento para o anonimato. O cálculo da perda de informações em um tratamento para anonimização de um conjunto de dados é da maior importância, principalmente para poder selecionar a solução ideal. Definir a métrica a ser usada depende de vários fatores, incluindo o tipo dos dados envolvidos, o algoritmo usado para obter privacidade e o possível uso final

dos dados. A distorção de dados é inevitável durante a anonimização e muitas métricas de qualidade são propostas para medir a qualidade de dados anônimos. Entretanto, a maioria das métricas de perda de informações existentes são limitadas ao refletir a distorção dos dados, sem considerar aspectos semânticos e ontológicos. A anonimização diminui a utilidade da informação, portanto, a utilidade dos dados anônimos validados com base nestas métricas é restrita. A decisão pelo anonimato pode causar a perda de informações ao ponto que inviabilizam determinados objetivos na análise de dados. Com isto é fundamental medir a perda da informação através de cálculos estatísticos apropriados e documentados conforme preconiza o Statistical Disclosure Control (SDC) proposto Willembourg em 2012 . (SWEENEY, 2002b; WILLENBORG; DE WAAL, 2012)(MEASURING UTILITY AND INFORMATION LOSS — SDC PRACTICE GUIDE DOCUMENTATION, *[s. d.]*)

Uma medida útil para comparar um mesmo conjunto de dados antes e depois do tratamento para anonimização é comparar o número de valores ausentes antes e após o tratamento. Os valores ausentes são geralmente introduzidos após a supressão e indicam um maior grau de perda de informações. Outra estatística útil é o número de registros alterados em cada variável. Estas alterações podem ser contadas de maneira semelhante aos valores ausentes e incluem a supressão e a relocação de dados categóricos. Para comparar as variáveis contínuas, medidas tradicionais como média, covariância e correlação podem identificar mudanças indesejadas, pois estas métricas não devem apresentar diferenças. (SWEENEY; CROSAS; BAR-SINAI, *[s. d.]*)

Além da perda de informação (*information loss*), para quantificar a utilidade dos dados, existem métricas de qualidade como, a discernibilidade e o erro de reconstrução. A discernibilidade mede a cardinalidade entre classes e considera apenas o número de registros na classe equivalente, estimando a qualidade dos dados com base no tamanho das classes de

equivalência no conjunto de dados de saída. Assim, não captura a perda de informações causada pela generalização. É um contraponto a métrica de perda de informação, pois pode medir a cardinalidade da classe equivalente e a perda de informações, considerando a veracidade dos dados. Já a métrica do erro de reconstrução mede a semelhança entre o registro original e o registro anônimo e pode refletir a perda de informações e a veracidade dos dados. (LIU; WANG; FENG, 2010)

A information loss foi relatada por Kohlmayer et al. para quatro medidas de utilidade diferentes (Entropia Não Uniforme, Precisão, Discernibilidade e AECS-Average Equivalence Class Size) em cinco conjuntos de dados diferentes. Os valores demonstraram a perda de informações como uma porcentagem ao anonimizar dados com generalização e supressão utilizando um limite de supressão de 5%. A utilidade do conjunto de dados com supressão aumentou em até 39% quando medida com a Entropia Não Uniforme (k-anonymity), 61% quando medido com Precisão (k-anonymity), 35% quando medido com Discernibilidade (k-anonymity e o-presença) e 89% quando medido com AECS (l-diversidade e o-presença). Em resumo, os resultados destacam o valor da supressão para anonimizar dados com saída de alta qualidade. (KOHLMAYER; PRASSER; KUHN, 2015).

Como extensão metodológica, técnicas que demandam maior elaboração podem comparar os dados reais e quantificar a distância entre o conjunto de dados original X e o conjunto de dados tratado Z através da métrica IL1s, onde X e Z contêm apenas variáveis contínuas que medem as distâncias de uma estruturação sintática das variáveis originais. (MATEO-SANZ; SEBÉ; DOMINGO-FERRER, 2004)

Quanto mais específico a utilidade de um algoritmo, menor é a sua viabilidade. Entretanto, técnicas híbridas e que somam estes conceitos são cada vez mais comuns. A Apple divulgou sua estratégia para coletar dados anônimos dos seus usuários com a

finalidade de aprimorar seus produtos e para isto, divulgou que utiliza um algoritmo proprietário e híbrido, que implementa privacidade diferencial em diferentes níveis, sendo o coeficiente “E” variável entre 6 e 10 para dados de suporte aos usuários e de até 43 para formulação de bases de testes. Já o Google tornou seu algoritmo de anonimização aberto em 2019. Chamado de RAPPOR (Randomized Aggregatable Privacy-Preserving Ordinal Response), o produto é considerado uma ferramenta estatística para recolher dados estatísticos de usuários de programas de computador e vem sendo desenvolvido desde 2014. Ele implementa a privacidade diferencial no resultado das pesquisas realizadas no buscador do Google e também no navegador Chrome, onde desempenha um papel vital para avaliar as vulnerabilidades dos usuários na internet. Apesar do uso restrito, tudo é feito através de uma biblioteca de funções transparentes e que tem o código aberto ao público. (ERLINGSSON; PIHUR; KOROLOVA, 2014; GREENBERG *et al.*, 2016)

Na medida em que a ciência aprofunda os conceitos sobre ataques, adversários, probabilidades e outras brechas que podem revelar a identidade das pessoas em dados considerados desidentificados, torna-se imprescindível o estudo e desenvolvimento desta área de pesquisa que trata de um conceito basal no estabelecimento da sociedade moderna que é o direito à privacidade (SWEENEY, 2002b, 2015).

Concluímos, que o anonimato é desafiador e as estratégias de anonimização e desidentificação ainda são limitadas, pois nem todas as necessidades de uso de um S-RES podem ser antecipadas e catalogadas. Logo, os tipos de dados invariavelmente precisam ser estatisticamente avaliados de forma objetiva antes e depois do tratamento para anonimização e uso na pesquisa, levando em consideração cada tipo de necessidade estabelecida pelos diferentes tipos de projetos. (KUSHIDA *et al.*, 2012; SWEENEY, 1997; YE; CHEN, 2011)

4. OBJETIVOS

Objetivo Geral

- Disponibilizar os meios para a utilização de conjuntos de dados anonimizados em modelos estatísticos e em aplicações com IA na saúde.

Objetivos Específicos

- Desenvolver métodos computacionais de tratamento de dados para privacidade dos pacientes.

- Verificar se a base original e a base anonimizada mantém as mesmas propriedades estatísticas.

ARTIGOS

1 - Ontologia para Anonimização de Dados Hospitalares

(Vaz T.A., Dora J.M., Lamb L.C., Camey S.A.)

Resumo

Este artigo apresenta o desenvolvimento de uma nova ontologia de domínio no escopo da epidemiologia, da medicina, da estatística e da ciência da computação. Para isto, utilizamos a terminologia definida pela legislação vigente e representamos o tratamento sistemático de dados hospitalares preparados com o anonimato para o uso em aplicações de Inteligência Artificial (IA) na saúde. Utilizamos um método que consiste em 7 passos pragmáticos para construção de uma nova ontologia, incluindo: a definição de escopo, a seleção de conhecimento, a revisão de termos importantes, a elaboração das classes que descrevem delineamentos utilizados em estudos epidemiológicos, os paradigmas de aprendizado de máquina, os tipos de dados e atributos, tipos de riscos aos quais os dados preparados podem estar expostos, tipos de ataques que podem acontecer contra a privacidade, técnicas de preparo para mitigar a reidentificação, modelos de privacidade que podem ser adotados e métricas para medir os efeitos da anonimização. Ao final, demonstramos como utilizar estas definições de acordo com as suas propriedades e relações, criando uma instância da ontologia para exemplificar como a anonimização pode ser implementada para o desenvolvimento e validação da IA nos hospitais.

Introdução

No âmbito das pesquisas realizadas com bases de dados contendo registros hospitalares, os dados considerados anônimos podem não apresentar as propriedades estatísticas necessárias para garantir o anonimato (SWEENEY, 2015). Isto é um problema para os pesquisadores da saúde que utilizam dados para o desenvolvimento de Inteligência Artificial (IA), entre outros estudos centrados na análise de dados (ROCHER; HENDRICKX; DE MONTJOYE, 2019). Para fazer o uso destes dados na pesquisa em saúde precisamos compreender como tratá-los preservando a privacidade das pessoas envolvidas mitigando o risco de re-identificação de acordo com a legislação vigente (SPENGLER; PRASSER, 2019).

No mundo, as nações possuem leis de proteção de dados que determinam os mecanismos que precisam ser adotados para que os trabalhos desenvolvidos utilizando dados pessoais e dados pessoais sensíveis sejam conduzidos. Com estas leis, cresce a necessidade de segurança da informação, governança, educação, auditoria e fiscalização do tratamento de dados, o que faz os custos inerentes ao uso dos dados aumentarem (ROCHER; HENDRICKX; DE MONTJOYE, 2019; SULLIVAN, 2004). Entretanto, algumas leis de privacidade, incluindo a Lei Geral de Proteção de Dados (LGPD) brasileira e a General Data Protection Regulation (GDPR) adotada pela União Europeia, definem que os dados quando têm ou adquirem propriedades de anonimato, não são mais objetos de regulamentação (BRASIL, 2018; UE, 2016). A LGPD define um dado anonimizado como: “dado relativo a titular que não possa ser identificado, considerando a utilização de meios técnicos razoáveis e disponíveis na ocasião de seu tratamento” (BRASIL, 2018). A GDPR define a anonimização como “informações que não se relacionam com uma pessoa singular identificada ou identificável ou a outros dados pessoais tornados anônimos". (UE, 2016)

Com a premência da anonimização dos dados hospitalares também surgem oportunidades na medida em que cada vez mais precisamos compartilhar com outros pesquisadores os dados analisados. Isto é feito para promover a reprodutibilidade dos experimentos, dar transparência nos métodos utilizados e nos resultados alcançados e os dados anônimos permitem isto (MARK, 2016). Mas para tratar este discurso de forma lógica, torna-se necessário representar semanticamente o quê dados de registros hospitalares anonimizados realmente são através de uma ontologia específica (ABHYANKAR; DEMNER-FUSHMAN; MCDONALD, 2012; GRUBER, 1993; REPS *et al.*, 2018).

Este trabalho apresenta a Ontologia de Registros Hospitalares Brasileiros (**ORHBR),** pronúncia *órber*. Serve para conectar conceitos multidisciplinares sobre privacidade e ciências de dados oriundos da epidemiologia, medicina, estatística e da ciências da computação, formando a compreensão das estruturas necessárias para descrever o pensamento sobre a anonimização de dados hospitalares.

Metodologia

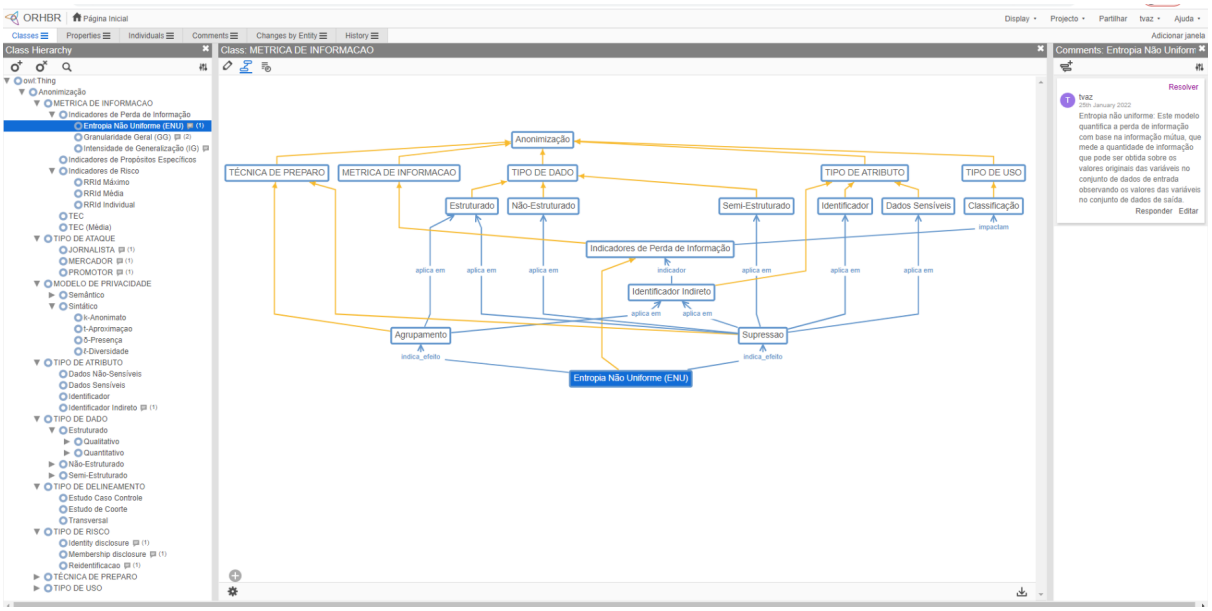
Para desenvolver a ORHBR utilizamos a metodologia proposta pelos pesquisadores NOY et al. (2001) para a construção de novas ontologias em 7 etapas. Na etapa 1 realizamos a definição de domínio e escopo, destacando o que não faz parte do escopo. Na etapa 2 acontece a seleção do conhecimento, incluindo a revisão de ontologias

semelhantes e a definição da proposta de extensão e adaptação. Na etapa 3 são listados os termos importantes do escopo definido, apresentando em formato de revisão de literatura dos dicionários de terminologias e outras padronizações da área da saúde, mas que não se relacionam a este tema específico. Nas etapas 4, 5 e 6 acontece o processo de trabalho para a criação de classes, propriedades e relações, definindo uma a uma em uma ferramenta especializada para criação de ontologias. O último passo é a etapa 7, quando definimos uma instância da ontologia (NOY; MCGUINNESS; OTHERS, 2001).

Desenvolvimento

Para desenvolver a ORBHR seguindo a metodologia proposta, executamos as seguintes etapas com o suporte da ferramenta *protege.standford.edu* (Figura 1) disponível na internet. *(NOY; MCGUINNESS; OTHERS, 2001)*

Figura 1 - Tela do aplicativo WebProtégé mostrando a estrutura de classes com um exemplo de propriedades e relações da ontologia ORHBR.



Fonte: Elaborado pelos autores, 2022.

1. Definição de domínio e escopo

O *escopo* inclui o necessário para avaliar se a anonimização de um determinado conjunto de dados hospitalares (o que consideramos uma instância desta ontologia) está sendo representada semanticamente de forma adequada.

Definimos *anonimização* como o preparo de um conjunto de dados para impossibilitar a reidentificação do titular dos dados, considerando a utilização de meios técnicos razoáveis e disponíveis na ocasião de seu tratamento. (BRASIL, 2018) Entretanto, existem outros métodos que não são considerados métodos de anonimização e por isso ficam fora do escopo deste trabalho, incluindo: a desidentificação através da remoção de dados predefinidos para dificultar o vínculo entre o titular e seus dados (SULLIVAN, 2004), a pseudonimização através da substituição dos identificadores do titular por uma chave alternativa (OLIVEIRA; MADEIRA; MONTEIRO, 2020) e o uso de criptografia (incluindo: função *hash*, *blockchain* e outras) para tornar o conjunto de dados secretos, mas que podem ser descriptografados e identificados. (KRAWIEC *et al.*, 2016)

Esta ontologia pretende apoiar os pesquisadores a responder perguntas qualitativas para caracterização dos diferentes estudos observacionais clínicos e epidemiológicos que usam registros hospitalares anonimizados.

Não fazem parte do escopo desta ontologia o escopo da saúde dos pacientes, pois não se trata de um novo vocabulário de termos médicos. A ORHBR também não é uma ontologia para organizar a informação dos sistemas informatizados de prontuário eletrônico, pois se trata de uma ontologia de domínio específico dos métodos de anonimização para pesquisas que utilizem IA em hospitais. A ORHBR foi construída de forma agnóstica, para ser utilizada independente de outros padrões e convenções que possam estar sendo utilizados pelos pesquisadores.

2. Seleção do Conhecimento

Utilizamos como referência o trabalho de QUEIROZ (2016) que apresenta uma ontologia de domínio para preservação de privacidade em dados publicados pelo governo federal brasileiro para fins de controle de acesso à informação, propondo as principais classes necessárias para tratar o tema (QUEIROZ; LINO; GUSTAVO H M, 2016). BATET et al. (2011) estabeleceram uma ontologia que define métricas para computar a similaridade semântica na biomedicina, importantes para definir resultados comparáveis neste tema (BATET; SÁNCHEZ; VALLS, 2011). LLUIS (2011) apresenta, em sua tese de doutorado, uma ontologia para as propriedades estatísticas da anonimização com e sem perturbação dos dados, propondo definições dos tipos de tratamento de dados possíveis e dos tipos de ameaça às quais os dados estão expostos (MARTÍNEZ LLUÍS; OTHERS, [*s. d.*]). PANOV et al. (2016) propõem uma ontologia genérica de tipos de dados chamada OntoDT, estabelecendo o conceito de tipagem de dados (PANOV; SOLDATOVA; DŽEROSKI, 2016).

A proposta do ORHBR nasce no escopo da LGPD - Lei Geral de Proteção de Dados (BRASIL, 2018), enquanto a referência selecionada para este trabalho teve sua origem objetivando a proteção de documentos públicos do governo federal fornecidos ao público em geral no âmbito da lei brasileira que regula o acesso à informação (BRASIL, [*s. d.*]) .

3. Termos importantes

De acordo com a legislação vigente no Brasil, *dado pessoal* é um termo que define se a informação está relacionada a uma pessoa natural identificada ou identificável. Dado pessoal sensível é o dado genético ou biométrico de uma pessoa, informação sobre a saúde, origem racial ou étnica, filiações à organização de caráter religioso, filosófico ou político. O termo *tratamento de dados*, define toda operação realizada com dados pessoais, incluindo coleta, recepção, classificação, acesso aos dados, todas formas de processamento, armazenamento em diferentes mídias e o descarte (BRASIL, 2018). *Identificadores* são todos os dados que permitem apontar diretamente sem o uso de outras informações quem é o titular de um dado. *Identificadores indiretos* são dados sobre um paciente que podem ser encontrados em outras fontes de informações públicas.

Os seguintes termos usuais são utilizados para designar dados identificadores e identificadores indiretos: nome completo, primeiro nome, sobrenome, endereço (todas as subdivisões geográficas menores que o estado, incluindo endereço, município e CEP), todos os elementos de datas relacionadas a um indivíduo (incluindo data de nascimento, data de admissão, data de alta, data de falecimento e idade exata, contatos (números de telefone, fax, endereço de e-mail, redes sociais), código de identificação (segurança social, plano de saúde, número do prontuário médico, conta bancária, cartão de crédito, certificados, notas fiscais e números de série dos produtos hospitalares e dispositivos médicos), URL, cookie, número IP, nome de usuário, identificadores biométricos (digital de dedo, retina ou voz), imagens e sons das pessoas (incluindo imagens de medicina diagnóstica - não se limitando às imagens fotográficas do rosto), amostras biológicas coletadas das pessoas e armazenadas em biobancos e serviços de medicina personalizada, com potencial de identificar através do DNA, ou outro método, a origem do material (SULLIVAN, 2004).

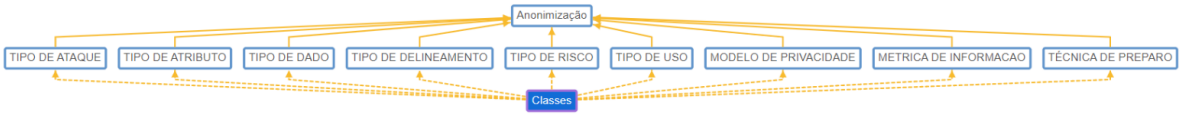
Os demais termos importantes para compreensão e uso desta ontologia estão estabelecidos em padrões internacionais e nos dicionários. Utilizamos como referência para este trabalho o dicionário da epidemiologia escrito pelo Prof. Miguel Porta, feito com o patrocínio da Associação Internacional de Epidemiologia (PORTA, 2014). Na medicina

utilizamos os conceitos da SNOMED-CT, que determina padrões internacionais para os termos médicos. Na ciência da computação adotamos os termos propostos por KERR (2016) e na informática em saúde utilizamos o dicionário de termos, acrônimos e organizações da saúde publicados pela Health Informatics and Management Systems Society (HIMSS).

4. Criação de classes

As *classes* representam as estruturas centrais de uma ontologia e foram definidas de acordo com as necessidades identificadas durante a análise de requisitos para criar a primeira instância. Na Figura 2 são apresentadas as nove classes definidas.

Figura 2: Classes da Ontologia ORHBR com conceitos multidisciplinares.



Fonte: Elaborado pelos autores, 2022.

A seguir, apresentamos a descrição de cada uma das classes, conceituando os valores que pertencem a cada uma de suas propriedades.

Tipo de Delineamento

Dentro do escopo desta ontologia, identificamos 3 tipos de delineamentos utilizados em estudos de pesquisa epidemiológica que podem ser realizados usando registros secundários anonimizados. Para detalhamentos sobre os principais delineamentos epidemiológicos indicaremos referências (NUNES; CAMEY; GUIMARÃES, 2013).

- **Transversal:** estudo observacional realizado examinando um conjunto de dados em um determinado momento do tempo para estimar a frequência de um determinado evento. A exposição e o desfecho são coletados da base de dados sem a informação de datas associadas aos eventos. Servem para analisar a prevalência e a associação entre exposição (expostos e não

expostos) e a classe de desfecho, que é usualmente binária, mas pode apresentar diferentes tipos de dados. (PORTA, 2014)

- **Caso-controle:** estudo observacional realizado para comparar pessoas com um desfecho de interesse positivo e um grupo controle de pessoas que não apresentam o desfecho. Usualmente utiliza longos períodos de coleta de dados de forma sistematizada. Permite identificar fatores de risco em doenças raras, estudar a etiologia das doenças e a análise da razão de chances. (PORTA, 2014)

- **Coorte**: estudo observacional realizado para acompanhar um grupo de pessoas ao longo do tempo, para avaliar os riscos e benefícios do uso de determinada intervenção ou medicamento e estudar a evolução e o prognóstico das doenças. Compara a incidência de uma doença na população usando uma proporção (risco relativo) ou uma diferença (risco atribuível).

Estes tipos de estudos possuem finalidades e limitações que precisam ser abordadas pelos(as) cientistas da computação durante experimentos com IA aplicados na saúde. Como por exemplo: os dados de saúde observados em um S-RES não registram a integralidade dos eventos de saúde do paciente e demandam análises específicas para contornar possíveis vieses de pesquisa (PORTA, 2014). Além disso, estes dados estão sujeitos a diferentes tipos de erros sistemáticos que refletem a qualidade como os funcionários de um hospital utilizam o S-RES.

Os demais delineamentos de estudos epidemiológicos, em geral, não permitem o uso da anonimização de forma sistemática. Os relatos de casos fazem a descrição detalhada de casos isolados, e não envolve uma análise centrada em dados. Os estudos experimentais (por exemplo: ensaios clínicos randomizados) implicam no acompanhamento prospectivo dos sujeitos de pesquisa. Mesmo que sejam tomadas medidas para preservação da privacidade utilizando o preparo de dados, por questões de segurança relacionada a saúde dos participantes, a maioria destes estudos precisa ter formas de reidentificar os titulares dos dados, tornando a anonimização uma alternativa inviável.

A partir do delineamento de um estudo, podemos identificar uma ou mais relações existentes com as demais classes apresentadas a seguir.

Tipo de Dado

No âmbito da anonimização, utilizamos a expressão *conjunto de dados* (tradução livre para *datasets*), para designar as estruturas contendo linhas e colunas de dados, também chamadas de tabelas, planilhas ou dados tabulares. Chamaremos de variáveis todas as colunas, também chamadas de covariáveis ou *features*. Para a compreensão dos tipos de dados existentes em um registro hospitalar, classificamos os tipos de dados em subclasses para representar diferentes tópicos.

- **Natureza:** Qualitativo (ordinal e nominal), Quantitativo (discreto e contínuo). (WILLENBORG; DE WAAL, 2012)
- **Estrutura:** Estruturado, semi-estruturado, não-estruturado. (KIMBALL; CASERTA, 2011)
- **Computação:** *int*, *float*, *char* (exemplos da linguagem C++) (IBM, 1993).
- **Metadados:** Taxonomia (*labels*) e dicionários. (CHUTE *et al.*, 2010)
- **Formato Digital do Conjunto de Dados:** Texto-plano (UTF-8, ISO-9071 e outros), Proprietário, Criptografado. (IBM, 1993)
- **Localização:** Português-BR, Inglês-EUA e outras, define a língua, fuso horário e valores de referência que podem variar conforme o local. (IBM, 1993)

Apresentados os tipos de dados que podem ser utilizados, podemos agora classificar uma variável quanto a sua perspectiva para anonimização e que chamaremos de tipos de atributos.

Tipo de Atributo

Os atributos definem como cada uma das variáveis será tratada durante a anonimização da informação.

- **Identificadores**: dados que vinculam diretamente a pessoa natural que é a titular dos dados. Inclui, mas não limita-se ao nome completo, prontuário do paciente, número dos documentos pessoais, registro de licença profissional, entre outros, que precisam ser removidos durante o tratamento para anonimização. (BRASIL, 2018)

- **Identificadores indiretos**: quando combinados podem revelar a identidade de dados pessoais, mas podem ser tratados com um algoritmo de privacidade para implementar o k-anonimato. (BRASIL, 2018; SAMARATI; SWEENEY, [*s. d.*])
- **Dados Pessoais Sensíveis**: todas as demais informações sobre a saúde de uma pessoa. Para fins de anonimização, podem ser tratadas com os algoritmos l-diversidade e t-aproximação. (MACHANAVAJJHALA *et al.*, 2007)

A partir das definições dos tipos de atributos, podemos analisar os tipos de risco, os ataques e os modelos de privacidade que podem ser utilizados para mitigar os riscos contra a privacidade.

Tipos de Risco

Os três principais riscos relacionados à privacidade aos quais os conjuntos de dados estão ameaçados são:

- **Divulgação de identidade (reidentificação):** é um risco de alto impacto na privacidade, pois quando um invasor tem sucesso na reidentificação ele aprende todas as informações confidenciais sobre o titular dos dados contidas no conjunto. (SWEENEY, 2002)

- **Divulgação de atributos:** é um risco de médio impacto na privacidade, pois quando um invasor tem sucesso é divulgado apenas o valor de algumas variáveis do conjunto, que podem permitir inferir quais são os titulares dos dados contidos no conjunto, mas não apontar quem é quem. (PRASSER *et al.*, 2016)

- **Divulgação de associação:** é um risco de menor impacto pois o invasor não divulga diretamente nenhuma informação do próprio conjunto de dados, mas permite determinar se o titular está ou não dentro do conjunto de dados. (PRASSER *et al.*, 2016)

Tipo de Ataque

Existem três modelos de ataques que podem ser utilizados para tentar identificar dados considerados anônimos.

- **Modelo Jornalista**: ataca para divulgar a identidade de um titular dos dados em específico. Utiliza a técnica de *data linkage* para relacionar identificadores indiretos contidos no conjunto de dados com outras informações públicas existentes na internet (SWEENEY, 2015).

- **Modelo Promotor**: ataca para divulgar a identidade do titular dos dados ou de um atributo específico, utilizando como conhecimento prévio ou *background knowledge* se os dados de interesse estão ou não contidos no conjunto de dados. (PRASSER *et al.*, 2016)

- **Modelo Mercador**: quando não existe um alvo específico, mas visa identificar um grande número de titulares dos dados existentes em um conjunto. (PRASSER *et al.*, 2016)

Modelo de Privacidade

Um registro eletrônico hospitalar pode ter os riscos de reidentificação mitigados com um modelo de privacidade: (AGGARWAL; YU, 2008)

- **k-Anonimato:** utiliza mudanças de agrupamento e supressão de identificadores indiretos para garantir que os dados de um indivíduo são indistinguíveis de k-1 outros indivíduos. (SWEENEY, 2002).

- **l-Diversidade**, é uma extensão ao k-anonimato na medida que utiliza a mesma proteção dada aos identificadores indiretos também aos dados pessoais sensíveis do conjunto, aumentando a complexidade computacional na medida em que novas categorias precisam ser definidas para cada valor

existente nos conjuntos de dados da saúde. (MACHANAVAJJHALA *et al.*, 2007)

- **t-Aproximação:** propõe superar a limitação do l-diversidade redefinindo a distribuição dos valores de cada variável ao invés do agrupamento. (RAJENDRAN; JAYABALAN; RANA, 2017)

- **δ-Presença:** pode ser usado para proteger os dados da divulgação de membros (*membership disclosure*), onde um conjunto de dados revela a probabilidade de um indivíduo da população estar contido no conjunto de dados (RAJENDRAN; JAYABALAN; RANA, 2017).

Existem outros modelos que adaptam os métodos apresentados e estes não serão tratados neste trabalho. (KHAN; FOLEY; O'SULLIVAN, 2021). Com a definição dos modelos de privacidade, podemos aplicar as técnicas de preparo relacionadas.

Técnica de Preparo

O ato determinante para preservar a privacidade de um conjunto de dados é o seu preparo com uma ou mais técnica específicas utilizadas de acordo com o modelo de privacidade selecionado:

- **Supressão**: Eliminação dos dados que possam identificar indiretamente uma pessoa. Para isto, são utilizados algoritmos que podem suprimir integralmente ou mascarar parcialmente as observações, variáveis ou valores específicos dentro de um conjunto de dados (SAMARATI; SWEENEY, [*s. d.*])

- **Agrupamento**: Classificação dos pacientes em categorias (também chamado de *generalization)*. O agrupamento de valores, permite que todos os pacientes que possuem a mesma categoria sejam classificados nos mesmos agrupamentos. (SAMARATI; SWEENEY, [*s. d.*])

- **Perturbação**: Inclusão de ruído ou sujeira, modificando os dados de forma intencional para dificultar a re-identificação. É utilizado para implementar em grandes bases de dados o método de privacidade diferencial. Devido ao tratamento que utiliza a modificação dos dados reais, não trataremos destes modelos durante este trabalho (DWORK, 2008).

Métrica de Informação

Ao utilizar as técnicas de preparo, temos o efeito da diminuição do risco de reidentificação e também da perda da utilidade da informação. (SWEENEY, 2002; WILLENBORG; DE WAAL, 2012)(MEASURING UTILITY AND INFORMATION LOSS — SDC PRACTICE GUIDE DOCUMENTATION, [*s. d.*]). Em geral, estas métricas podem utilizar tanto os dados originais (entrada) quanto os dados anonimizados (saída) para serem computadas.

- **Tamanho de Equivalência da Classe (TEC):** é a quantidade de indivíduos dentro do conjunto que possuem os mesmos identificadores indiretos. (SWEENEY, 2002)

- **Entropia Não Uniforme (ENU):** compara a diferença antes e depois da anonimização do tamanho de equivalência das classes em todo o conjunto de dados e individualmente para cada um dos atributos. (PRASSER; BILD; KUHN, 2016)

- **Risco de Reidentificação Individual (RRId Individual):** informa o risco de quebra de privacidade de um determinado titular dos dados a partir do TEC do registro. (SWEENEY, 2002)

- **Risco Máximo e Médio de Reidentificação (RRId Máximo e RRId Médio):** é o valor máximo e a média do RRId Individual em todo conjunto de dados. (LEFEVRE; DEWITT; RAMAKRISHNAN, 2006)

- **Intensidade de Generalização (IG):** identifica a perda de informação entre o conjunto original e o conjunto anonimizado a partir do somatório da quantidade de valores que foram modificados durante a anonimização (SWEENEY, 2002).

- **Granularidade Geral (GG):** compara a quantidade distinta de valores existentes em uma variável antes e depois da anonimização para mostrar a perda de informação. (IYENGAR, 2002)

- **Método de propósito específico:** compara modelos derivados de diferentes formas de preparo de um mesmo conjunto de dados. (PRASSER; BILD; KUHN, 2016)

Tipo de Uso

A anonimização de um conjunto de dados impacta nos indicadores de perda de informação e consequentemente nos resultados da utilização das técnicas de IA. Logo, o preparo precisa ser feito de acordo com estes indicadores e considerando os diferentes tipos de uso, conforme proposto por LEFEVRE et. al (2008):

- **Análise de regressão linear:** envolve encontrar um modelo linear que descreve ou prevê o valor de uma variável dependente quantitativa em função das demais variáveis existentes no conjunto. As análises de regressão linear podem ser implementadas com IA utilizando o aprendizado de máquina supervisionado, incluindo: regressão linear, redes neurais, árvores de regressão, entre outros. (JAMES *et al.*, 2014)

- **Classificação:** é a atribuição de variáveis qualitativas que representam classes com valores pré-determinados (também chamados de alvos, categorias, variáveis dependente, *targets* ou *labels*) utilizando um procedimento sistemático com base nas variáveis observadas. É uma tarefa realizada em estudos transversais e tem como característica não utilizar informações de datas relacionadas aos dados. Os algoritmos classificadores de IA podem utilizar diferentes abordagens de aprendizado, incluindo:

supervisionado, semi-supervisionado e não supervisionado. Os principais métodos para classificação incluem: regressão logística, métodos baseados em árvores de decisão, redes neurais, *linear discriminant analysis*, *clustering*, *boosting* e *support vector machines* (SVM).

- **Recuperação de Informação:** uma seleção (ou *query*) envolve um conjunto de critérios utilizados para filtrar os dados e definir grupos para uma população (subpopulação). Combinações utilizando operadores lógicos possibilitam a formulação de *queries* complexas geralmente implementadas com a linguagem *Structured Query Languages* (SQL). O uso do processamento de linguagem natural (PLN) e outras técnicas que utilizam a IA permitem a seleção de informações também no texto livre e em imagens, por exemplo. Uma tarefa de seleção pode ser necessária durante diferentes momentos de um estudo. (LEFEVRE; DEWITT; RAMAKRISHNAN, 2008)

- **Clusterização:** envolve reconhecer, diferenciar e compreender como os dados podem ser agrupados em categorias. O agrupamento também é um tipo de preparo utilizado para implementar a anonimização e por isto dados agrupados também são muitas vezes considerados dados anónimos. Entretanto, algumas tarefas derivadas do agrupamento e que podem ser realizadas pela IA com demandas específicas para a anonimização são a classificação de tópicos, atribuição de taxonomias e o agrupamento em *clusters* (que pode ser feito utilizando o aprendizado de máquina não-supervisionado) (IBM, 1993)**.**

Existem muitas formas de utilizar um conjunto de dados anonimizado, focamos naquelas que utilizam o Aprendizado de Máquina (AM) para desenvolver algoritmos que tenham a capacidade de aprender com suas experiências e assim melhorar seu desempenho em determinadas tarefas necessárias para resolver problemas. (LAMPROPOULOS; TSIHRINTZIS, [*s. d.*]) Estas tarefas podem ser implementadas com IA em diferentes programas de computador e também via programação utilizando bibliotecas e pacotes de aprendizado de máquina Scikit Learn e Stats para as linguagens Python e R, respectivamente. (KUHN, 2008; LORENZONI *et al.*, 2019; PEDREGOSA *et al.*, 2011; SANCHES, 2003)

5. Criação de propriedades

Com a definição das classes e suas subclasses, podemos criar as propriedades que indicam as relações existentes entre as classes, respondendo perguntas utilizando termos importantes para anonimização de registros hospitalares, como: Qual o tipo de risco que o conjunto de dados está exposto? Qual o tipo de dados dos atributos que são identificadores indiretos no conjunto? Quais modelos de privacidade diminuem o risco contra os tipos de ataque previstos? Quais os riscos de reidentificação do conjunto em caso de um determinado tipo de ataque? Quais métricas podem ser utilizadas para identificar a perda de informação em dados preparados com supressão? Respondendo estas perguntas, podemos definir as propriedades como funções que conectam as classes:

- **tem risco (Tipo de Delineamento, Tipo de Risco):** define que um determinado tipo de delineamento utilizado em estudos epidemiológicos apresenta determinados tipos de riscos para privacidade, por exemplo:

    - *Transversal <tem risco> Reidentificação*

- **tem impacto (Tipo de Atributo, Tipo de Risco):** define que um determinado tipo de atributo do conjunto apresenta determinados tipos de riscos para privacidade, por exemplo:

    - *Identificador Indireto <tem impacto> Reidentificação*

- **faz ataque (Tipo de Ataque, Tipo de Risco):** define que um determinado tipo de ataque pode ser feito colocando em risco a privacidade, por exemplo:

    - *Jornalista <faz ataque> Reidentificação*

- **mitiga risco (Tipo de Ataque, Modelo de Privacidade):** define que um determinado tipo de ataque pode ser mitigado utilizando um determinado modelo de privacidade, por exemplo:

    - *Jornalista <mitiga risco> k-Anonimato*

- **faz tratamento (Modelo de Privacidade, Técnica de Preparo):** define que um determinado modelo de privacidade pode utilizar determinadas técnicas de preparo dos dados na sua implementação, por exemplo:

    - *k-Anonimato <faz tratamento> {Supressão, Agrupamento}*

- **faz preparo (Técnica de Preparo, Tipo de Dados):** define que técnicas de preparo podem ser utilizadas nos tipos de dados, por exemplo:

    - *Agrupamento <faz preparo> {Qualitativo, Quantitativo}*

- **tem tipo (Tipo de Dados, {Subclasses}):** define um determinado tipo de dados, de acordo com uma determinada lista de subclasses, por exemplo:

    - *Natureza <tem tipo> {Quantitativo,Qualitativo}*
    - *Quantitativo <tem tipo> {Contínuo, Discreto}*
    - *Qualitativo <tem tipo> {Nominal, Ordinal}*

- **medir informação (Tipo de Dados, Métrica de Informação):** define para um tipo de dados, qual a métrica apresenta a perda de informação durante a anonimização:

    - *Tipo de Dados Nominal <medir informação> ENU*

- **identifica impacto (Métrica de Informação, Tipo de Tarefa):** define qual tipo de tarefa realizada pela IA pode ter o impacto da anonimização avaliado por uma métrica:

    - *ENU <identifica impacto> Classificação*

Definidas as propriedades das classes podemos definir as relações existentes entre elas com um indivíduo (ou instância) no mundo real.

6. Criação das relações

Indivíduos (ou objetos) na vida real podem utilizar os termos desta ontologia e assim criar as relações necessárias para tratar a anonimização de acordo com a finalidade do uso

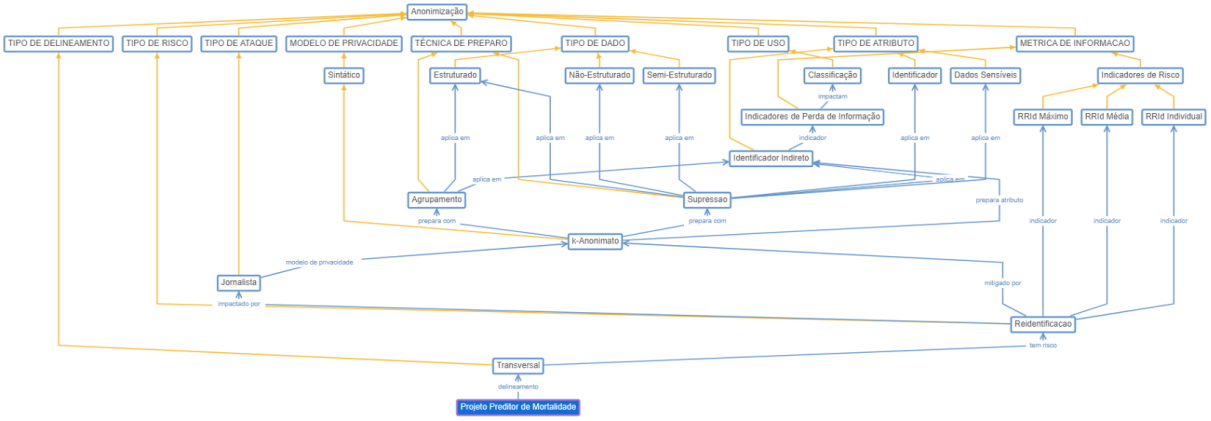
que os conjuntos de dados terão. (NOY; MCGUINNESS; OTHERS, 2001). Vamos utilizar como exemplo, um pesquisador que vai estudar como desenvolver uma aplicação de IA para predizer a mortalidade hospitalar, classificando quais pacientes têm o maior risco de ir a óbito nas primeiras horas de internação.

A primeira pergunta a ser feita para iniciar esta relação entre o pesquisador e os métodos de anonimização, deve ser feita para compreender a sua pesquisa e assim identificar suas propriedades na ontologia. Para iniciar esta conversa, podemos perguntar: Qual é o delineamento do estudo? A partir desta resposta, utilizamos as propriedades que conectam as classes e as propriedades no domínio da anonimização para definir uma sequência de relacionamentos para tratar o assunto no contexto de um estudo específico, e assim podemos definir uma instância da ontologia.

7. Criação de Instância

As instâncias representam aplicações reais da ontologia (NOY; MCGUINNESS; OTHERS, 2001). A figura 3 exemplifica uma instância. A partir da ontologia é possível definir um projeto hipotético para ilustrar o seu funcionamento. A nova instância consiste em: 1) um estudo transversal sobre mortalidade hospitalar, 2) a técnica deIA utilizada para classificar pacientes com risco de óbito 3) os tipos de dados utilizados são estruturados, de natureza qualitativa e quantitativa, 4) os tipos de dados computacionais da natureza qualitativa são *strings* e da natureza quantitativa são *float*, 5) a variavel prontuário é um atributo identificador direto, 6) as variaveis atributos idade e sexo são atributos identificadores indiretos, 7) objetivo é mitigar o risco de reidentificação, 8) defendendo de ataques do tipo jornalista, 9) adotando o modelo de privacidade k-anonymity, 10) utilizando técnicas de preparo de supressão e agrupamento, 11) com o risco de reidentificação mensurado pelas métricas do RRid, RRid Médio e RRid Máximo.

Figura 3 - Exemplo de instância da ORHBR.



Fonte: Elaborado pelos autores, 2022.

Discussão

Em sua revisão sobre o tema, CHEVRIER et. al (2019) tratam da variabilidade observada nos termos que definem os métodos de preparo e na forma como os termos desidentificação e anonimização são utilizados, enfatizando a necessidade de definições objetivas centradas na legislação para aprimorar a educação e disseminação de informações sobre o assunto.(CHEVRIER *et al.*, 2019) Esta ontologia define uma estrutura formal para compreender o risco de reidentificação de registros hospitalares, muitas vezes traduzindo diferentes conceitos entre diferentes áreas e isto exigiu fazer escolhas. Nossa expectativa não está em mudar a cultura das áreas de pesquisa consolidadas, mas sim, ajudar a formar uma nova comunidade de pesquisadores que vão utilizar conjuntos de dados hospitalares não só para treinar e usar IA, mas para potencializar o uso de dados em diferentes tipos de estudos epidemiológicos, na gestão hospitalar e na inovação em saúde como um todo.

Existem previsões legais para o uso de dados hospitalares e uma delas é o consentimento informado do paciente, autorizando o uso de seus dados na pesquisa, por exemplo (BRASIL, 2018). Para obter o consentimento informado dos pacientes, na prática é preciso interagir com eles. A forma como este consentimento é obtido pode introduzir mais de um tipo de viés na pesquisa clínica (EMAM *et al.*, 2013). A anonimização de conjuntos de dados na primeira oportunidade após a coleta é uma alternativa para não precisar obter este consentimento, entretanto o tratamento de dados para anonimização pode interferir na utilidade dos dados e não permitir atingir os objetivos de uma pesquisa. Assim, a ORHBR

descreve também a semântica para compreender e justificar porque uma determinada pesquisa precisa trabalhar com dados identificados.

Nosso estudo foi limitado na elaboração de apenas uma instância, utilizada para exemplificar o uso da ontologia. Entendemos que este é o início de um novo campo na pesquisa em saúde e a criação de outros tipos de instância podem demandar novos requisitos. Acreditamos que uma ontologia como a ORHBR pode ser adotada pela autoridade nacional de proteção de dados pessoais e depositada no Repositório de Vocabulários e Ontologias do Governo Eletrônico, potencializando o uso por outros pesquisadores e a manutenção da ontologia, visando estender suas classes incluindo novos tipos de dados (por exemplo: imagem e som) e tipos de delineamento (por exemplo: estudos contrafactuais), apoiando a manutenção de uma política de anonimização que potencializa o uso dos dados.

Conclusão

Definimos uma ontologia para alavancar a cultura da privacidade na pesquisa com registros hospitalares, sem perder de vista as oportunidades de resolver problemas de diferentes tipos com o uso da IA. A partir da adoção da **ORHBR** pesquisadores de diferentes áreas poderão compartilhar e reutilizar o conhecimento técnico sobre a anonimização de registros hospitalares utilizando um mesmo vocabulário. Com o uso desta ontologia, esperamos poder comparar quantitativamente e qualitativamente diferentes modelos de privacidade que informam os riscos e a perda da informação durante o processo de anonimização.

Referências

ABHYANKAR, Swapna; DEMNER-FUSHMAN, Dina; MCDONALD, Clement J. Standardizing clinical laboratory data for secondary use. **Journal of biomedical informatics**, [*s. l.*], v. 45, n. 4, p. 642–650, 2012.

AGGARWAL, Charu C.; YU, Philip S. A General Survey of Privacy-Preserving Data Mining Models and Algorithms. *Em*: AGGARWAL, Charu C.; YU, Philip S. (org.). **Privacy-Preserving Data Mining: Models and Algorithms**. Boston, MA: Springer US, 2008. p. 11–52.

BATET, Montserrat; SÁNCHEZ, David; VALLS, Aida. An ontology-based measure to

compute semantic similarity in biomedicine. **Journal of biomedical informatics**, [*s. l.*], v. 44, n. 1, p. 118–125, 2011.

BRASIL. **LAI**. [*S. l.*], [*s. d.*]. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm. Acesso em: 20 abr. 2022.

BRASIL. **LGPD: A Lei Geral de Proteção de Dados Pessoais**. [*S. l.*], 2018. Disponível em: http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2018/Lei/L13709.htm. Acesso em: 11 fev. 2019.

CHEVRIER, Raphaël *et al.* Use and Understanding of Anonymization and De-Identification in the Biomedical Literature: Scoping Review. **Journal of medical Internet research**, [*s. l.*], v. 21, n. 5, p. e13484, 2019.

CHUTE, Christopher G. *et al.* The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. **Journal of the American Medical Informatics Association: JAMIA**, [*s. l.*], v. 17, n. 2, p. 131–135, 2010.

DWORK, Cynthia. Differential Privacy: A Survey of Results. *Em*: , 2008. **Theory and Applications of Models of Computation**. [*S. l.*]: Springer Berlin Heidelberg, 2008. p. 1–19.

EMAM, Khaled El *et al.* A Review of Evidence on Consent Bias in Research. **The American journal of bioethics: AJOB**, [*s. l.*], v. 13, n. 4, p. 42–44, 2013.

GENERAL DATA PROTECTION REGULATION (GDPR) – OFFICIAL LEGAL TEXT. [*S. l.*], 2016. Disponível em: https://gdpr-info.eu/. Acesso em: 20 abr. 2022.

GRUBER, Thomas R. A translation approach to portable ontology specifications. **Knowledge Acquisition**, [*s. l.*], v. 5, n. 2, p. 199–220, 1993.

IBM. **IBM Dictionary of Computing**. 10th. ed. USA: McGraw-Hill, Inc., 1993.

IYENGAR, Vijay S. Transforming data to satisfy privacy constraints. *Em*: , 2002, Edmonton, Alberta, Canada. **Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining**. New York, NY, USA: Association for Computing Machinery, 2002. p. 279–288. Disponível em: Acesso em: 21 abr. 2022.

JAMES, Gareth *et al.* **An Introduction to Statistical Learning: with Applications in R**. 1st ed. 2013, Corr. 7th printing 2017 editioned. [*S. l.*]: Springer New York, 2014.

KHAN, Muhammad; FOLEY, Simon; O'SULLIVAN, Barry. From k-anonymity to Differential Privacy: A Brief Introduction to Formal Privacy Models. [*s. l.*], 2021. Disponível em: https://hal.archives-ouvertes.fr/hal-03226881/.

KIMBALL, Ralph; CASERTA, Joe. **The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data**. [*S. l.*]: John Wiley & Sons, 2011.

KRAWIEC, R. J. *et al.* Blockchain: Opportunities for health care. *Em*: , 2016. **Proc. NIST Workshop Blockchain Healthcare**. [*S. l.: s. n.*], 2016. p. 1–16.

KUHN, Max. Building Predictive Models in R Using the caret Package. **Journal of**

statistical software, [*s. l.*], v. 28, p. 1–26, 2008. Disponível em: Acesso em: 24 abr. 2022.

LAMPROPOULOS, Aristomenis S.; TSIHRINTZIS, George A. **Machine Learning Paradigms**. [*S. l.*]: Springer International Publishing, [*s. d.*]. *E-book*. Disponível em: Acesso em: 1 jun. 2022.

LEFEVRE, K.; DEWITT, D. J.; RAMAKRISHNAN, R. Mondrian Multidimensional K-Anonymity. *Em*: , 2006. **22nd International Conference on Data Engineering (ICDE'06)**. [*S. l.: s. n.*], 2006. p. 25–25.

LEFEVRE, Kristen; DEWITT, David J.; RAMAKRISHNAN, Raghu. Workload-aware anonymization techniques for large-scale datasets. **ACM Trans. Database Syst.**, New York, NY, USA, v. 33, n. 3, p. 1–47, 2008.

LORENZONI, Giulia *et al.* Comparison of Machine Learning Techniques for Prediction of Hospitalization in Heart Failure Patients. **Journal of clinical medicine research**, [*s. l.*], v. 8, n. 9, 2019. Disponível em: http://dx.doi.org/10.3390/jcm8091298.

MACHANAVAJJHALA, Ashwin *et al. L*-diversity: Privacy beyond *k*-anonymity. **ACM transactions on knowledge discovery from data**, New York, NY, USA, v. 1, n. 1, p. 3 – es, 2007.

MARK, Roger. The Story of MIMIC. *Em*: MIT CRITICAL DATA (org.). **Secondary Analysis of Electronic Health Records**. Cham: Springer International Publishing, 2016. p. 43–49.

MARTÍNEZ LLUÍS, Sergio; OTHERS. **Ontology based semantic anonymisation of microdata**. [*s. d.*]. - Universitat Rovira i Virgili, [*s. l.*], [*s. d.*]. Disponível em: https://www.tdx.cat/handle/10803/108961.

MEASURING UTILITY AND INFORMATION LOSS — SDC PRACTICE GUIDE DOCUMENTATION. [*S. l.*], [*s. d.*]. Disponível em: https://sdcpractice.readthedocs.io/en/latest/utility.html. Acesso em: 4 dez. 2019.

NOY, Natalya F.; MCGUINNESS, Deborah L.; OTHERS. **Ontology development 101: A guide to creating your first ontology**. [*S. l.*]: Stanford knowledge systems laboratory technical report KSL-01-05 and …, 2001. Disponível em: https://protege.stanford.edu/publications/ontology_development/ontology101.pdf.

NUNES, L. N.; CAMEY, S. A.; GUIMARÃES, L. S. P. Os principais delineamentos na Epidemiologia. **Vol. 33, no. 2 (2013), p …**, [*s. l.*], 2013. Disponível em: https://www.lume.ufrgs.br/handle/10183/158317.

OLIVEIRA, Erick de; MADEIRA, Humberto dos Santos; MONTEIRO, Pedro Augusto Migliari. A lei geral de proteção de dados pessoais e a anonimização de dados: uma aplicação da técnica em uma base de dados real. [*s. l.*], 2020. Disponível em: http://ric.cps.sp.gov.br/handle/123456789/5258. Acesso em: 27 abr. 2022.

PANOV, Panče; SOLDATOVA, Larisa N.; DŽEROSKI, Sašo. Generic ontology of datatypes. **Information sciences**, [*s. l.*], v. 329, p. 900–920, 2016.

PEDREGOSA, Fabian *et al.* Scikit-learn: Machine learning in Python. **the Journal of machine Learning research**, [*s. l.*], v. 12, p. 2825–2830, 2011.

PORTA, Miquel. **A Dictionary of Epidemiology**. [*S. l.*]: Oxford, 2014.

PRASSER, Fabian *et al.* Lightning: Utility-Driven Anonymization of High-Dimensional Data. **Transactions on Data Privacy. Foundations and Technologies**, [*s. l.*], v. 9, n. 2, p. 161–185, 2016.

PRASSER, Fabian; BILD, Raffael; KUHN, Klaus A. A Generic Method for Assessing the Quality of De-Identified Health Data. **Studies in health technology and informatics**, [*s. l.*], 2016. Disponível em: https://www.researchgate.net/publication/317006066_A_Generic_Method_for_Assessing_the_Quality_of_De-Identified_Health_Data. Acesso em: 19 maio 2022.

QUEIROZ, Maria J.; LINO, Natasha C. Q.; GUSTAVO H M. Uma Ontologia de Domínio para Preservação de Privacidade em Dados Publicados pelo Governo Brasileiro. *Em*: , 2016. **Anais do XII Simpósio Brasileiro de Sistemas de Informação**. [*S. l.*]: SBC, 2016. p. 009–016. Disponível em: Acesso em: 2 nov. 2020.

RAJENDRAN, Keerthana; JAYABALAN, Manoj; RANA, Muhammad Ehsan. A Study on k-anonymity, l-diversity, and t-closeness Techniques focusing Medical Data. [*s. l.*], v. 17, n. 12, 2017. Disponível em: http://dx.doi.org/. Acesso em: 6 dez. 2019.

REPS, Jenna M. *et al.* Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. **Journal of the American Medical Informatics Association: JAMIA**, [*s. l.*], v. 25, n. 8, p. 969–975, 2018.

ROCHER, Luc; HENDRICKX, Julien M.; DE MONTJOYE, Yves-Alexandre. Estimating the success of re-identifications in incomplete datasets using generative models. **Nature communications**, [*s. l.*], v. 10, n. 1, p. 3069, 2019.

SAMARATI, Pierangela; SWEENEY, Latanya. **Protecting privacy when disclosing information: K-anonymity and its enforcement through generalization and suppression**. [*S. l.*],.2003. Disponível em: https://dataprivacylab.org/dataprivacy/projects/kanonymity/paper3.pdf. Acesso em: 22 abr. 2022.

SANCHES, Marcelo Kaminski. Aprendizado de maquina semi-supervisionado: proposta de um algoritmo para rotular exemplos a partir de poucos exemplos rotulados. [*s. l.*], Disponível em: https://teses.usp.br/teses/disponiveis/55/55134/tde-12102003-140536/publico/Dissertacao_MKS.pdf.

SPENGLER, Helmut; PRASSER, Fabian. Protecting Biomedical Data Against Attribute Disclosure. **Studies in health technology and informatics**, [*s. l.*], v. 267, p. 207–214, 2019.

SULLIVAN, June M. **HIPAA: A Practical Guide to the Privacy and Security of Health Data**. [*S. l.*]: American Bar Association, 2004.

SWEENEY, Latanya. ACHIEVING k-ANONYMITY PRIVACY PROTECTION USING GENERALIZATION AND SUPPRESSION. **International Journal of Uncertainty, Fuzziness and Knowlege-Based Systems**, [*s. l.*], v. 10, n. 05, p. 571–588, 2002.

SWEENEY, Latanya. Only you, your doctor, and many others may know. **Technology**

Science, [*s. l.*], v. 2015092903, n. 9, p. 29, 2015.

WILLENBORG, Leon; DE WAAL, Ton. **Elements of Statistical Disclosure Control**. [*S. l.*]: Springer Science & Business Media, 2012.

2 - BRHIM - Preparação De Dados Hospitalares para Estudos Observacionais

Autores: Vaz T.A., Dora J.M., Lamb L.C., Camey S.A.

RESUMO

Importância: Os riscos de reidentificação dos dados hospitalares é alto e devido às leis de privacidade e aos avanços da Inteligência Artificial (IA) existe uma demanda por métodos de preparo para anonimização de registros secundários.

Objetivos: Definir um processo para preparar dados utilizando diferentes métodos que mitigam riscos de reidentificação dos pacientes em bases de registros hospitalares e comparar os efeitos destes em uma aplicação de IA.

Materiais e Métodos: A nova receita foi definida com 5 etapas: organização, análise descritiva, pseudonimização, desidentificação e anonimização. Utilizamos 2 anos de dados hospitalares (30.464 internações) para aplicar a receita e formar quatro conjuntos: dados brutos, dados pseudonimizados, dados desidentificados e dados anonimizados. Definimos as métricas de risco de reidentificação e perda da informação nos diferentes conjuntos. Realizamos um evento *datathon* para desenvolver uma IA preditora de mortalidade hospitalar, que após foi replicada nos diferentes conjuntos de dados para comparar os efeitos do preparo.

Resultados: O algoritmo *k-anonymity* com o valor de k=20 reduziu o risco médio de reidentificação de 11,5% no conjunto original para 2,2% no conjunto anonimizado. Ao mesmo tempo, foi feita a remoção de 436 registros (1,4%) no conjunto anônimo para atingir este resultado. Treinamos a IA com os dados originais, desidentificados e anonimizados obtendo *Area Under the Curve* - Receiver Operating Characteristic Curve (AUC-ROC) iguais a 86,2%, 85,7% e 85,5% respectivamente.

Conclusão e Relevância: Comparando os resultados da IA utilizando dados originais e dados anonimizados, tivemos uma diferença inferior a 1% na AUC-ROC, ao mesmo tempo em que o risco máximo de reidentificação de um paciente utilizando seus identificadores indiretos foi reduzido em 95% utilizando a k-anonimato. Esta receita de preparo pode ser sistematizada em outras instituições e assim cada pesquisador pode definir os seus parâmetros de privacidade para estabelecer quais são os tipos de riscos e os limites aceitos na anonimização, bem como torná-los transparentes.

INTRODUÇÃO

Um estudo observacional pode ser feito a partir de dados que são coletados para um objetivo definido pelo pesquisador (considerados registros primários) ou dados coletados para outra finalidade, mas que podem ser usados pelo pesquisador para seu objetivo (considerados registros secundários). No caso de um hospital que utiliza um Sistema de Registros Eletrônicos em Saúde (S-RES), os dados são coletados primariamente com a finalidade de alimentar um prontuário eletrônico e possibilitar a assistência aos pacientes. Então dizemos que estes bancos de dados são registros secundários quando utilizados para pesquisa (SORENSEN; SABROE; OLSEN, 1996).

Nestes registros secundários os dados não são capturados em um formato definido a priori pelo pesquisador e isto agrega complexidade no seu preparo para análise, desenvolvimento e validação de novos estudos. Outra diferença importante é que ao utilizar registros secundários pode não haver consentimento do paciente para uso dos dados em pesquisas e isto também adiciona complexidade no processo. Um exemplo de uso dos registros secundários preparados é quando os pesquisadores utilizam um S-RES como fonte de dados para desenvolver e validar uma Inteligência Artificial (IA) capaz de predizer o risco de morte de um paciente durante sua internação no hospital (MARK, 2016).

Em geral, executar uma receita de preparo destes registros secundários pode tornar-se um processo lento, sujeito a repetições e caro quando realizado sob a demanda específica de cada pesquisador (CHUTE *et al.*, 2010). O preparo dos dados precisa contemplar requisitos específicos de cada estudo que podem ser definidos seguindo um percurso claro, incluindo etapas que devem ser cumpridas para contemplar as leis de proteção de dados pessoais (LGPD) e contribuindo com os demais pesquisadores que desejam avaliar se os dados de um centro estão ou não devidamente preparados para iniciar o desenvolvimento de modelos, incluindo aqueles que utilizam IA (BOONSTRA; VERSLUIS; VOS, 2014; HYPPÖNEN *et al.*, 2014; WANG *et al.*, 2016).

Este trabalho constrói uma metodologia de preparo de dados incluindo a anonimização, em até 5 etapas, mede a utilidade da informação tratada, identificando através de métricas os riscos de privacidade, e compara os modelos obtidos pelas técnicas de IA na base original e na base anonimizada chamada Base de Registro Hospitalar para Informações e Metadados (BRHIM).

MÉTODO

Trata-se de um estudo observacional em base de registros secundários. Utilizamos 2 anos de dados do Hospital de Clínicas de Porto Alegre (HCPA) que usa o S-RES chamado Sistema de Gestão em Saúde (AGHUse). O sistema é desenvolvido e mantido no HCPA com diferentes módulos de negócio utilizados pelos profissionais da saúde, que geram e mantém as informações para assistência, gestão administrativa, pesquisa clínica, entre outros. (HCPA, 2014; SILVA, 2016; VAZ, 2017).

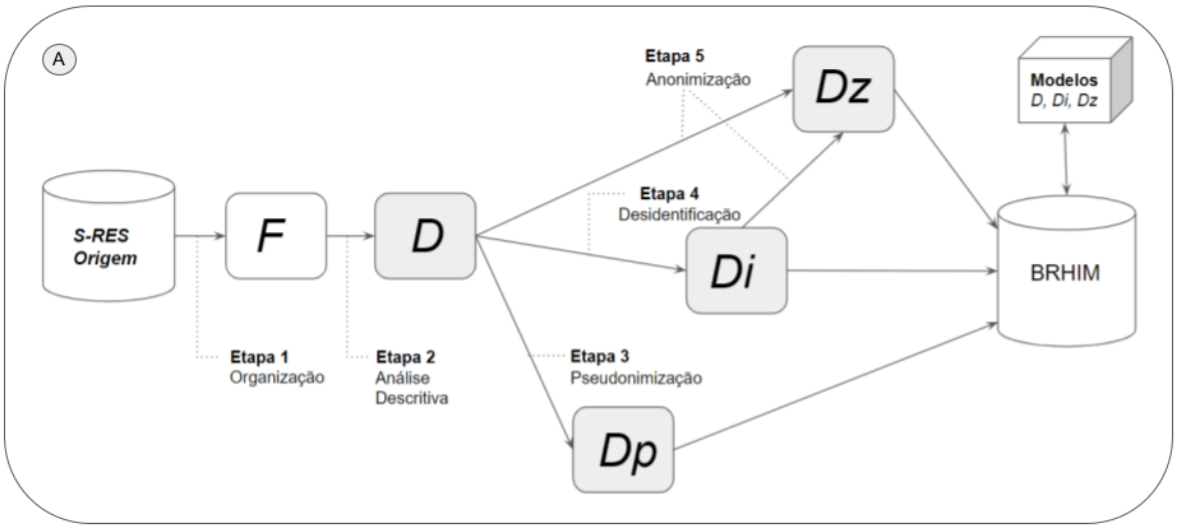
Os dados foram extraídos com o uso de comandos em *Structured Query Language* (SQL) executados diretamente no Sistema de Gerenciamento de Banco de Dados (SGBD) do AGHUse. Após, utilizamos as linguagens Python e R Markdown na plataforma de ciência de dados Dataiku versão 9.1 (DATAIKU, [*s. d.*]; PEDREGOSA *et al.*, 2011; WICKHAM; GROLEMUND, 2016) em conjunto com a biblioteca ARX versão 3.9 (EICHER *et al.*, 2020) para realizar o preparo de quatro conjuntos de dados: *D* - Dados brutos , *Dp* - Dados pseudonimizados, *Di* - Dados desidentificados e *Dz* - Dados anonimizados. Ao final, aplicamos uma IA preditora de mortalidade hospitalar possibilitando comparar as métricas dos resultados obtidos utilizando os diferentes conjuntos (BRAJER *et al.*, 2020).

Para desenvolver cada consulta SQL e depois cada uma das etapas da receita utilizamos o método Planejar-Desenvolver-Controlar-Agir (PDCA) (SHEWHART, 1940). Utilizamos um painel para classificar e visualizar cada tarefa de preparo com cores indicativas das etapas, possibilitando as definições que integram a receita de preparo apresentada a seguir.

Receita de Preparo

O preparo dos dados é apresentado em 5 diferentes etapas. Entre a ETAPA 1 e a ETAPA 2 incluímos a formatação e a análise descritiva que é comum em todos conjuntos de dados. A ETAPA 3 prepara uma base pseudonimizada. A ETAPA 4 é baseada em regras para desidentificar a base de dados. E a ETAPA 5 faz o tratamento dos dados para anonimização. As etapas de 3 a 5 são independentes e são utilizadas conforme a necessidade de cada caso, sendo que as etapas 4 e 5 podem ser realizadas também em conjunto. A Figura 1 apresenta dados fictícios para demonstrar o fluxo dos conjuntos de dados durante o preparo.

Figura 1 - (A) Fluxo do preparo de conjuntos original, pseudonimizado, desidentificado e anonimizado. (B) Exemplos destes conjuntos.



B

D: Conjunto original

Prontuário*	Sexo**	Idade**	Risco de Re-Identificação
1001	F	85	100%
1002	F	85	100%
1003	F	90	100%
1004	F	81	100%
1005	M	75	100%
1006	M	75	100%
1007	M	75	100%
1008	M	75	100%

* Identificador direto ** Identificador indireto

Dp: Conjunto pseudonimizado

Id ***	Sexo	Idade	Risco de Re-Identificação
2hj3f2	F	85	100%
g59dk	F	85	100%
9fh4o	F	90	100%
2j4pk	F	81	100%
1g7gf	M	75	100%
a6sf2	M	75	100%
ne38f	M	75	100%
68xi9	M	75	100%

*** Código Criptografado

Di: Conjunto desidentificado

Prontuário	Sexo	Idade	Risco de Re-Identificação
*	F	85	50%
*	F	85	50%
*	F	>89	100%
*	F	81	100%
*	M	75	25%
*	M	75	25%
*	M	75	25%
*	M	75	25%

Dz: Conjunto anonimizado

Prontuário	Sexo	Idade	Risco de Re-Identificação
*	F	>80	25%
*	F	>80	25%
*	F	>80	25%
*	F	>80	25%
*	M	75	25%
*	M	75	25%
*	M	75	25%
*	M	75	25%

Fonte: Elaborado pelos autores, 2022.

Antes do Preparo

Grande parte do trabalho e do retrabalho de preparação dos dados acontece, ou pode ser evitado, no momento em que elaboramos as *queries* escritas em SQL utilizadas no processo *Extract-Transform-Load* (ETL) (KIMBALL; CASERTA, 2011). Para qualificar este trabalho é possível transformar os dados extraídos pelas *queries* para um modelo de dados clínicos. Selecionamos o modelo proposto por HRIPCSAK et al. (2015) e utilizamos os seus conceitos para construir a base secundária, vinculando através de identificadores únicos as tabelas Pacientes, Período Observado, Observação, Localização, Medicamento, Visita, Procedimento, Diagnóstico, Desfecho e Notas da Equipe Assistencial (HRIPCSAK *et al.*, 2015).

Antes de iniciar o preparo revisamos os filtros de cada uma das 10 tabelas. Por exemplo: verificando se o resultado não inclui datas fora dos intervalos de datas solicitadas, ou se existem pacientes com idade diferente das especificadas, ou outras inconsistências com os critérios de

inclusão e exclusão da pesquisa. Após, deve-se verificar se as colunas solicitadas estão presentes e por fim, se as colunas contêm dados.

Se os filtros estiverem errados, ou uma variável obrigatória estiver ausente, ou contiver apenas dados ausentes, a *query* utilizada no ETL precisa ser revisada junto com especialistas e quando necessário a extração dos dados precisa ser refeita antes de iniciar o preparo.

ETAPA 1 - Organização

A formatação dos dados é a primeira etapa do preparo. Segundo WICKHAM (2014), 80% do tempo total de uma análise de dados é gasto no processo de preparo dos dados e existe pouca pesquisa em como formatar dados eficientemente. Para realizá-la, primeiro organizamos o trabalho reunindo todos os dados e dicionários envolvidos em um só local, identificando as características de cada uma das tabelas e colunas com o apoio de analistas e especialistas na assistência e na gestão do hospital (KAHN *et al.*, 2016). Com a compreensão dos dados que vão ser preparados, podemos iniciar as tarefas de preparo de forma ordenada:

A. Definir **convenções** para dar nomes aos conjuntos de dados e as variáveis. Algumas ferramentas modernas que utilizam sistemas de codificação configuráveis (exemplo: LATIN-1, UTF8, etc) permitem utilizar espaço em branco entre as palavras que compõem o nome das tabelas e das colunas, assim como manter todos os acentos da língua portuguesa. Isto é desejável pois permite desenvolver relatórios e painéis de dados sem precisar reescrever estes nomes. Por exemplo: ao invés de denominar uma coluna de ALTA_MEDICA, dar preferência para o nome “Alta Médica”. Entretanto, algumas ferramentas como o banco de dados MySQL e algumas versões de banco de dados proprietários não suportam o uso de caracteres especiais ou reservados e por isto esta definição deve ser tomada cautelosamente (REWATKAR; LANJEWAR, 2010).
B. Definir o **tipo computacional** das variáveis (inteiro, decimal, categoria, data ou texto, os tipos variam de acordo com a linguagem utilizada). Cada tipo de dado de uma linguagem de programação ocupa um determinado espaço na memória. Por exemplo, um atributo numérico inteiro pode ocupar 8 bits, enquanto um caractere para representar o mesmo número pode ocupar 32 bits. Isto impacta diretamente no tamanho final do conjunto de dados, logo utilizar os tipos de dados corretamente otimiza o uso da memória do computador (PANOV; SOLDATOVA; DŽEROSKI, 2016).
C. Combinar elementos de **data e hora**, incluindo o *timezone,* em uma só coluna do tipo UTC com o formato YYYY-MM-DDThh:mm:ss.sTZD especificado pela ISO 8601. Esta formatação garante a assertividade durante as conversões de formato de data necessárias para análise. (ORGANIZATION, 2004)

D. Ajustar os formatos dos números decimais para a localidade da língua portuguesa no Brasil (PT-BR) no formato NNN.NNN,NN.
E. Definição e aplicação de convenções para **tradução dos valores categóricos** (exemplo: ‘S’ ou ‘SIM’) para qualificar o entendimento do significado da informação.
F. Substituir ou excluir **caracteres indevidos** em variáveis numéricas (Por exemplo: >, <, E, M, $).
G. **Categorização dos valores numéricos** definidos a priori para a pesquisa. Exemplo: categorizar as idades em faixas etárias.
H. Criar **fórmulas** para calcular a conversão de unidades (por exemplo: centímetros transformados em metros) ou criação de novas variáveis (por exemplo: calcular IMC a partir das informações de peso e altura).
I. **Tratar o texto livre**, definir procedimentos para eliminar ou manter quebra de linha, acentos, caracteres especiais e etc. Para analisar a integridade da informação (se a remoção de quebras de linha e caracteres especiais alterou o significado do texto) é preciso identificar em uma amostra, se após a formatação ainda é possível a compreensão da sua leitura. Pode fazer parte desta etapa a construção de variáveis a partir do texto livre. Como por exemplo: a partir do texto da anamnese categorizar em uma nova variável se o paciente é fumante, ex-fumante ou nunca fumou. Isto é feito com o uso de expressões regulares, redes neurais, entre outras técnicas que não serão abordadas neste trabalho (SZLOSEK; FERRETT, 2016).

É possível detectar outras informações que precisam ser corrigidas ou eliminadas, isso é feito usando análise descritiva (ETAPA 2). Para isto, os dados precisam estar devidamente organizados no conjunto *F* (Figura 1) de acordo com um significado, algo conhecido na comunidade de desenvolvedores R como dados *tidy* (tradução livre: organizados) (WICKHAM; GROLEMUND, 2016):

J. Cada coluna de uma tabela é uma variável.
K. Cada linha é uma observação única (por exemplo: cada linha é um paciente).
L. Cada célula é uma única medida observada de uma variável.

ETAPA 2 - Análise Descritiva

Verificada a formatação inicial dos dados, em conjunto com especialistas (minimamente um(a) estatístico(a) e o pesquisador da área) é possível detectar informações que precisam ser corrigidas, eliminadas ou até mesmo demandar a interrupção do trabalho para retornar para antes do preparo (entre outras decisões possíveis) em caso de inconsistências. Isso é feito com a análise descritiva do conjunto, que pode utilizar medidas resumo e ferramentas de visualização de dados para realização das seguintes tarefas:

A. Analisar a **distribuição de frequência** separadamente de cada variável qualitativas para detectar respostas não previstas (exemplo: na variável sexo aparecer uma resposta igual a X) ou frequencias muito baixas ou muito altas que não eram previstas (exemplo: 95% dos pacientes atendidos no HCPA serem do sexo masculino).
B. Analisar as **medidas resumo** das variáveis quantitativas. Calcular mínimo, máximo, medidas de localização ou tendência central (média, mediana, moda) e medidas de variação ou dispersão (desvio padrão, variância e intervalo interquartílica). Esta análise serve para identificar valores discrepantes (exemplo: IMC = 1,4) e falta de variabilidade (intervalo interquartílico igual a zero)..
C. Analisar as células ***missing*** (vazias ou com informação errada) em todas as variáveis do conjunto. A forma de tratamento de valores *missing* é específica para cada tipo de estudo, podendo implicar na remoção das linhas e das colunas, ou na imputação dos dados .

As ações que serão tomadas a partir do resultado dessa análise devem ser discutidas com os especialistas. Por exemplo, quando for encontrando um valor discrepante é necessário tomar a decisão de mantê-lo ou excluí-lo da base. Estas análises podem ser feitas em todo o conjunto de dados (visão geral), com recorte temporal (anual, mensal, diária, etc) caso existam variáveis do tipo data no conjunto, ou em subpopulações específicas definidas pelos especialistas (por exemplo: analisar a distribuição das variáveis agrupadas por faixa etária). Concluídas as duas primeiras etapas, temos então a base original *D*, composta por A_p variáveis, com *p* variando de 1 até o número de variáveis na base *D*. Denotaremos, portanto, $D(A_1, \ldots, A_p)$ um conjunto com *p* variáveis.

ETAPA 3 - Pseudo Anonimização

A pseudo anonimização é o processo de proteger um dado pessoal substituindo identificadores diretos (por exemplo: prontuário) por uma chave criptografada. Assim, através dessa chave é possível utilizar uma senha para reidentificar o titular dos dados. Ela pode ser feita com o uso de diferentes algoritmos de criptografía. Por exemplo: o conjunto de dados *Dp* foi pseudo anonimizado, tendo a coluna Prontuário substituída pela coluna Id. A nova coluna contém a mesma informação porém criptografada com o algoritmo SHA-2, criado pela Agência Norte Americana de Segurança (NSA). Os passos necessários para proteger uma coluna de identificação com a pseudo anonimização são:

A. Selecionar uma coluna somente com **valores únicos** (por exemplo: Prontuário) capaz de identificar cada uma das linhas e que dará origem ao novo código criptografado.
B. Gerar uma **senha** numérica aleatória secreta.

C. Aplicar uma **função *hash*** criptográfica (por exemplo: SHA-512) sobre a coluna selecionada, utilizando o valor de uma segunda coluna (por exemplo: Idade) concatenada com uma senha numérica secreta para implementar a criptografia. Este processo é conhecido como *salt and pepper* (sal e pimenta) e é utilizado para gerar uma chave que pode descriptografar os dados e voltar a identificação do registro com o uso de uma senha. (BHONGE; AMBAT; CHANDAVARKAR, 2020)

Apesar do nome contra intuitivo, a pseudo anonimização não tem similaridade com a anonimização. A criptografia não diminui o risco de reidentificação dos dados justamente por utilizar uma chave de recuperação. Somente a desidentificação e a anonimização são capazes de diminuir os riscos de reidentificação.

ETAPA 4 - Desidentificação

Utilizamos o conceito de desidentificação da *Health Insurance Portability and Accountability Act* (HIPAA) com 18 regras que possibilitam a execução de uma receita de preparo objetiva relacionada ao tratamento de desidentificação via supressão dos dados considerados *Personal Health Information* (PHI). (ALSHUGRAN; DICHTER, 2014).

Para desidentificar dados estruturados, podemos suprimir os dados removendo as colunas que são PHI ou suprimir via substituição os valores originais por marcadores do tipo <NOME_REMOVIDO>. O uso de marcadores também deve ser utilizado para desidentificar os campos de textos (anamneses, evoluções, laudos etc) utilizando técnicas de PLN (Processamento de Linguagem Natural) ou outras técnicas capazes de realizar os seguintes passos:

A. Substituir **nomes** de: pacientes e familiares, equipes assistenciais e unidades de atendimento pelo marcador <NOME REMOVIDO>.
B. Substituir **datas** e períodos inferiores a um ano, pelo marcador <DATA_REMOVIDA>
C. Substituir **números** de telefone ou fax pelo marcador <TELEFONE_REMOVIDO>
D. Substituir dados **geográficos**, incluindo nomes de locais e estabelecimentos pelo marcador <LOCAL_REMOVIDO>
E. Substituir **códigos** e números de identificação como RG, CPF, PRONTUÁRIO, CRM, previdência social, plano de saúde, pelo marcador <CODIGO_REMOVIDO>
F. Substituir **endereço** de e-mail e URLs da web pelo marcador <LINK_REMOVIDO>
G. Substituir **identificadores** de notas fiscais, placas de veículos e números de série pelo marcador <IDENTIFICADOR_REMOVIDO>

Após aplicar estas etapas de desidentificação temos a base Di. A próxima etapa pode ser aplicada diretamente na base original D ou na base desidentificada Di, sendo esta uma decisão específica em cada projeto.

ETAPA 5 - Anonimização

A anonimização transforma os dados de *D* (ou Di) para *Dz* objetivando mitigar os riscos de reidentificação existentes em 3 modelos de ataque à privacidade: jornalista, promotor e mercador (PRASSER *et al.*, 2014). Cada modelo de ataque possui algoritmos específicos para preservação da privacidade, e cada algoritmo possui hiperparâmetros que precisam ser definidos antes da sua execução.

Utilizaremos as seguintes definições básicas:

Identificadores Indiretos: são todas as variáveis definidas pelo pesquisador como atributos que podem ser combinados para reidentificar um registro. Usualmente estes atributos podem ser encontrados publicamente, como: idade, sexo, estado civil, cor da pele e nível de escolaridade.

Classe de Equivalência: registros com os mesmos identificadores indiretos possuem a mesma classe de equivalência (E_j). Portanto, $D = E_1 \ U \ ... \ U \ E_j$, onde *j* é o número de classes de equivalência no conjunto D.

Tamanho de Equivalência das Classes (TEC): é a quantidade de registros *i* que pertencem a mesma classe de equivalência. Por exemplo: na Figura 1, os atributos do conjunto *D* considerados identificadores indiretos são sexo e idade (A = {sexo, idade}). Logo, os prontuários 1001 e 1002 pertencem à mesma classe de equivalência e o TEC(D,A,*i*) é igual a 2, i=1001, 1002.

Hiperparâmetro k: o valor *k* é definido a priori pelo pesquisador e é utilizado pelos algoritmos de anonimização para limitar o tamanho mínimo do TEC em um conjunto anonimizado. Quanto maior o valor de *k*, maior será a privacidade de um conjunto, enquanto *k*=1 significa que os registros não são anonimizados.

A partir destas definições, o preparo para anonimização dos dados demanda os seguintes passos:

A. Determinar os **objetivos do anonimato**: 1) Defender a privacidade contra o ataque de um jornalista que tem como alvo reidentificar um indivíduo específico utilizando a técnica de *data linkage* para vincular identificadores indiretos de um registro a outras informações disponíveis publicamente (SWEENEY, 2002). 2) Defender o ataque de um promotor que possui conhecimento prévio (*background knowledge)* de que os dados sobre o indivíduo específico estão contidos no conjunto de dados (PRASSER *et al.*, 2014). 3) Defender contra o ataque de um mercador que não tem como alvo um indivíduo específico, mas visa obter sucesso identificando um grande número de indivíduos (DANKAR; EL EMAM, 2010).
B. Selecionar o **tipo de algoritmo** de privacidade a partir do objetivo do anonimato. 1) k-anonimato serve para mitigar o risco de ataques do tipo jornalista realizando a supressão ou a agrupamento dos identificadores indiretos com o valor de TEC <k (SWEENEY, 2002). 2) O tipo de ataque promotor reage aos algoritmos l-diversity e

t-closeness, que fazem supressão e a categorização de dados que não são identificadores indiretos (MACHANAVAJJHALA *et al.*, 2007). 3) Para o ataque mercador pode ser utilizado o algoritmo de privacidade diferencial, que faz a perturbação dos dados para atingir o objetivo proposto (DWORK, 2008).

C. Definir os **hiperparâmetros** do algoritmo de anonimização selecionado. Além do *k*, existem outros hiperparâmetros que variam conforme o algoritmo selecionado e estão detalhados na documentação da biblioteca ARX (PRASSER *et al.*, 2014). Os hiperparâmetros precisam ser definidos para cada estudo e podem ser utilizadas diferentes técnicas de otimização para selecionar seus valores (LEFEVRE; DEWITT; RAMAKRISHNAN, 2008).

D. Definir **agrupamentos** para todas as variáveis que são identificadores indiretos. Por exemplo: agrupar idade em faixa etária. Isto será utilizado pelo algoritmo, que pode decidir em manter a idade original ou utilizar um dos agrupamentos definidos para não precisar excluir um registro com TEC <=k e assim diminuir a perda de utilidade durante a anonimização (PRASSER *et al.*, 2014).

E. **Reescalar** todas as colunas numéricas utilizando o algoritmo *min-max scaler* ou semelhante. Reescalar valores numéricos diminui as chances de reidentificação quando o ataque é feito por um promotor ou mercador (PRASSER *et al.*, 2020).

F. Em geral o **texto livre não pode ser anonimizado**, somente desidentificado, devido às numerosas limitações técnicas existentes. Nesta etapa devemos remover todas as variáveis contendo texto livre existentes no conjunto (CHEVRIER et al., 2019).

Com estas definições é possível rodar os algoritmos e mitigar os riscos de cada tipo de ataque através de transformações que implicam em perda de informação. Ao final desta etapa, teremos o conjunto original $D(A_1,\ldots,A_p)$ contendo os identificadores indiretos $A = A_1,\ldots,A_p$ suprimidos ou agrupados no conjunto anonimizado $Dz(Az_1,\ldots,Az_p)$ onde $TEC(Dz, Az, i) >= k$, para todo i. Assim, fica determinado o **Risco de Reidentificação Máximo (RRId Máximo)**, que é a chance de sucesso que um ataque contra a privacidade pode ter e reidentificar ao menos um dos titulares presentes em um conjunto de dados, sendo o $RRid\ Máximo(D) = \frac{1}{k} * 100$.

Para identificar os impactos e a efetividade desta transformação definimos as métricas que medem o efeito da anonimização.

Métricas para Medir o Efeito do Preparo com Relação à Privacidade

As métricas para medir os efeitos para desidentificação e para anonimização no preparo dos dados podem ser divididas em dois grupos. O primeiro grupo contém as métricas que visam medir o risco em relação à privacidade a partir do TEC, incluindo as métricas de **Risco de Reidentificação Individual (RRid Individual)** e o **Risco de Reidentificação Médio (RRId Médio)**.

O segundo grupo mede o efeito das modificações na perda de informação (ou *information loss*), permitindo uma análise de todo conjunto ou para cada uma das variáveis com as métricas de de **Intensidade de Generalização (IG)**, de **Granularidade Geral (GG)** e da **Entropia Não Uniforme dos Atributos (ENU)** antes e depois da desidentificação e da anonimização.

Ao longo do texto chamaremos esta métricas de indicadores de risco e de indicadores de perda de informação.

RRid Individual

É o risco de reidentificação individual de cada registro i no conjunto de dados D contendo os identificadores indiretos *A,* depende do valor do TEC e é calculado utilizando a fórmula:

$$RRid(D, A,\ i)\ =\ \frac{1}{TEC(D, A, i)} * 100 \quad (1)$$

Para garantir que o RRid individual de uma base anonimizada fosse sempre menor ou igual ao RRId Máximo, bastaria fazer a exclusão automática de registros onde $RRid(D, A,\ i) < 1/k$. Entretanto, para preservar a utilidade um algoritmo de anonimização pode agrupar e suprimir determinados valores dos indicadores indiretos para não precisar excluí-los no conjunto Dz, e ainda assim assim manter o $TEC(Dz,\ Az,\ i) >= k$, que implica que RRid individual é menor ou igual a RRId Máximo para todos os registros. (SWEENEY, 2002)

A partir do RRid individual é possível calcular o valor médio do risco para o conjunto D.

RRid Médio

É a média do risco de reidentificação de um conjunto D contendo *n* registros i. Seu valor depende do valor do RRid Individual e é calculado com a fórmula:

$$RRid\ Médio(D,\ A)\ =\ \frac{RRid\ (D, A, i) + \ldots + RRid\ (D\ A, n)}{n} * 100 \quad (2)$$

O RRId Médio de todo o conjunto é importante, pois os algoritmos de anonimização tendem a preservar o RRId Individual com valores próximos ao RRId Máximo, independente das técnicas utilizadas em seu preparo. Entretanto o RRId Médio tende a ser mais suscetível para identificar melhorias no método de preparo dos indicadores indiretos e assim apontar aprimoramentos no risco de preservação da privacidade de um conjunto. No exemplo *Dz* da Figura 1, o RRId Médio de todo o conjunto é igual a 50%. (LEFEVRE; DEWITT; RAMAKRISHNAN, 2006)

A técnica apropriada de aprimoramento da anonimização envolve tratar os dados até que o valor do RRId Individual de cada registro seja menor ou igual a 33% (para evitar que existam registros únicos ou em duplas) e após avaliar a redução do RRId Médio de todo o conjunto,

possibilitando identificar melhorias nas técnicas de preparo. (COMMITTEE ON STRATEGIES FOR RESPONSIBLE SHARING OF CLINICAL TRIAL DATA; BOARD ON HEALTH SCIENCES POLICY; INSTITUTE OF MEDICINE, 2015)

Intensidade de Generalização (IG)

Este indicador serve para identificar a perda de informação entre os atributos de um conjunto que passa por desidentificação ou anonimização. Utilizamos a Intensidade de Generalização computar a quantidade de valores que foram modificados no preparo, comparando a igualdade dos valores antes e depois em todas as linhas *i* e colunas *j* de ambos conjuntos, com a seguinte fórmula:

$$IG(D, DZ) = (1 - \frac{\sum_{j=1}^{p} \sum_{i=1}^{n} I(A_{ij} \neq Az_{ij})}{n.p}) * 100, se\ n >= 0\ e\ p >= 0 \quad (3)$$

Também utilizamos a Intensidade de Generalização para avaliar a quantidade de valores que foram modificados em um único atributo A_p, com a seguinte fórmula:

$$IG(D, A_p, DZ, Az_p) = (1 - \frac{\sum_{i=1}^{n} I(A_{pi} \neq Az_{pi})}{n}) * 100, se\ n >= 0\ e\ p >= 0 \quad (4)$$

Onde: *I* é a função indicadora, ou seja, quando $A_{ij} \neq Az_{ij}$ ela assume valor 1, caso contrário assume valor 0. n é o total de linhas, p é o total de atributos, D é o conjunto original e Dz é o conjunto anonimo.Portanto, *IG(D, Dz) = 1* se não há modificação dos identificadores indiretos. Quanto mais categorizações e supressões ocorrem, aumenta a perda de informação e o IG se aproxima de zero. Se todos os valores dos indicadores indiretos passam por supressão ou categorização, temos que *IG(D, Dz) = 0* (SWEENEY, 2002).

Granularidade Geral (GG)

Para comparar a perda de informação dos identificadores, calculamos a Granularidade Geral dos atributos com a seguinte fórmula:

$$GG(A_p, Az_p) = \frac{Q_{Az_p}}{Q_{A_p}} * 100 \quad (5)$$

Onde Q_{Az_n} é a quantidade distinta de valores existentes em *Az* após o preparo e Q_{A_n} é a quantidade distinta de valores antes do preparo. Quanto mais próximo o valor de GG for de 100%, menor será a perda de qualidade da informação de *A* durante a anonimização. Por exemplo: ao preparar o atributo $A_{idade} = 73$ que pertence ao intervalo $71 \leq A_{idade} \leq 80$ (contendo 10 valores distintintos) , transformados em Dz em um conjunto de intervalos do tipo [71,75],[76,80] (contendo 2 valores distintos), logo teremos o $GG(A_p, Az_p)$ = *2/10 = 0,2*. (IYENGAR, 2002)

Entropia Não Uniforme (ENU)

A ENU é semelhante a GG, porém calcula a perda de informação de um atributo A_p a partir do TEC, utilizando a seguinte fórmula:

$$ENU(D, A_p, Dz, Az_p) = (1 - (\sum_{i=1}^{n} - log \frac{TEC\ (D, A_p, i)}{TEC\ (Dz, Az_p, i)})) * 100 \quad (6)$$

Na fórmula assume-se que o quociente é sempre menor ou igual a 100%, pois o TEC de *i* só pode aumentar durante o preparo. Assim, o logaritmo negativo da razão é sempre um número positivo e a soma de todos eles mostra a entropia do atributo. Quanto mais próximo de 1 for o valor de ENU, menor é a perda de informação da variável durante a anonimização de *D para Dz*. (PRASSER; BILD; KUHN, 2016)

Para medir os efeitos da desidentificação e da anonimização comparamos a performance de três modelos desenvolvidos com Aprendizado de Máquina (AM) utilizando os conjuntos *D, Di e Dz* para o treinamento e validação e analisamos as diferenças na sensibilidade, especificidade, acurácia e a *Area Under the Curve - Receiver Operating Characteristic Curve* (AUC-ROC) dos resultados. As variáveis, o algoritmo e os hiperparâmetros utilizados foram definidos durante um evento denominado DIA DATATHON. Com apoio da Diretoria de Pesquisa (DIPE) do HCPA, do Instituto de Informática da UFRGS e da Fundação Médica do Rio Grande do Sul, foram organizadas 4 equipes diferentes com 10 participantes cada. Os pesquisadores da área foram convidados e preencheram o TCLE (Termo de consentimento livre e esclarecido). As equipes receberam acesso ao *software* Dataiku instalado no HCPA por onde acessaram os dados e puderam rodar diferentes tipos de algoritmos preditores, incluindo redes neurais, árvores de decisão e outros. Ao final, o modelo da equipe vencedora classificou corretamente todos os casos de óbito hospitalar existentes entre os registros do conjunto de dados da competição. As métricas obtidas foram AUC-ROC = 91,7%, Sensibilidade = 87,3%, Especificidade = 81,1% e Acurácia = 81,6%. Os detalhes do modelo estão no material suplementar. (KAWAMURA, 2002)

RESULTADOS

A tabela 1 mostra estatísticas descritivas do conjunto de dados preparados para o desenvolvimento do preditor de mortalidade utilizado nos três diferentes cenários. Identificamos que 436 (1,5%) internações do conjunto original (D) foram suprimidas durante o processo de anonimização, sendo 40 (1,6%) casos de óbitos. Para definir o RRid máximo igual a 5% a anonimização dos identificadores indiretos: sexo, cor da pele, estado cívil e escolaridade tiveram 1,4% dos valores suprimidos. Como podemos ver na tabela 1 os conjuntos *Dp* e *D* têm as mesmas medidas, pois a criptografia aplicada no identificador não altera as demais variáveis em *Dp*.

Tabela 1 - Distribuição (n, percentual, média e desvio padrão) dos atributos nos conjuntos

	Conjunto *D*	**Conjunto *Dp***	**Conjunto *Di***	**Conjunto *Dz***
Total de Internações (n)	30.464	30.464	30.464	30.028
Idade - Média (DP)	61(18,4)	61(18,4)	61(18,5)	63(17,9)
Sexo (n, %)				
- Feminino	16.260(53,4)	16.260(53,4)	16.260(53,4)	16.096(53,6)
- Masculino	14.204 (46,6)	14.204 (46,6)	14.204 (46,6)	13.932 (46,3)
- *Missing*	0 (0)	0 (0)	0 (0)	0 (0)
Cor da Pele				
- Branca	26.048(85,5)	26.048(85,5)	26.048(85,5)	25.900 (86,2)
- Preta	3.009 (9,9)	3.009 (9,9)	3.009 (9,9)	2.886 (9,6)
- Parda	1.365 (4,5)	1.365 (4,5)	1.365 (4,5)	1.242 (4,1)
- Outras	42 (0,1)	42 (0,1)	42 (0,1)	0 (0)
- *Missing*	0 (0)	0 (0)	0 (0%)	0 (0)
Estado Civil,				
- Casado	13.058 (43,6)	13.058 (43,6)	13.058 (43,6)	13.058 (43,4)
- Outros	17.168 (56,4)	17.168 (56,4)	17.168 (56,4)	16.970 (56,5)
- *Missing*	0 (0)	0 (0)	0 (0)	0 (0)
Escolaridade				
- 1º Grau Incompleto	12.787 (42.0)	12.787 (42.0)	12.787 (42.0)	12.787 (42.5)
- 1º Grau Completo	4.833 (15.9)	4.833 (15.9)	4.833 (15.9)	4.833 (16.0)
- 2º Grau Incompleto	1.616 (5.3)	1.616 (5.3)	1.616 (5.3)	1.386 (4.6)
- 2º Grau Completo	5.843(19.2)	5.843(19.2)	5.843(19.2)	5.843(19.4)
- Superior	2.046(6.7)	2.046(6.7)	2.046(6.7)	2.046(6.8)
- Superior Incompleto	1.002(3.3)	1.002(3.3)	1.002(3.3)	796(2.6)
- Nenhum	2.337 (7.6%)	2.337 (7.6%)	2.337 (7.6%)	2.337 (7.7%)
- *Missing*	0(0)	0(0)	0(0)	0(0)
Admissão via Emergência	16.035(52,6)	16.035(52,6)	16.035(52,6)	16.035(53,4)
Realizou Cirurgia	7.008(23)	7.008(23)	7.008(23)	6.934 (23,0)
Óbitos	2.422(7,9)	2.422(7,9)	2.422 (7,9)	2.382 (7,9)

Fonte: Elaborado pelos autores, 2022.

Também podemos analisar que os riscos de reidentificação diminuíram durante a anonimização, conforme apresentado na Tabela 2. Ao implementar o k-anonymity o risco máximo de reidentificação em *Dz* foi definido em 5%. O risco médio de reidentificação diminuiu de 11,52% em D para 2,2% em *Dz*. Entretanto, a supressão e o agrupamento reduziram o número total de grupos distintos em todo o conjunto de 3512 para 664, aumentando o tamanho máximo de TEC de 130 para 134, e os valores mínimos de TEC de 1 para 20. Para avaliar a perda de informação entre os conjuntos durante o preparo para anonimização, computamos o indicador IG que mostrou uma redução de 6.9%.

Tabela 2 - Estatísticas das classes de equivalência e de risco de reidentificação em *D* e *Dz*

Métrica	*D*	*Dz*
Classes de Equivalência (n)	3.512	664
TEC Mínimo (n)	1	20
TEC Máximo (n)	130	134
TEC Médio (n)	8,6	45,2
IG (%)	100	93.1
Risco Máximo de Reidentificação (%)	100	5
Risco Médio de Reidentificação (%)	11,5	2,2

TEC: Tamanho de Equivalência das Classes, IG: Intensidade de Generalização

Fonte: Elaborado pelos autores, 2022.

Ao analisarmos as métricas que medem o efeito da anonimização nos identificadores indiretos (Tabela 3), temos que apenas 1,4% dos registros foram integralmente suprimidos durante a k-anonimização, enquanto 7,6% dos registros tiveram a sua idade suprimida, revelando a importância de tratar este atributo para diminuir o risco de re-identificação dos pacientes.

Considerando o IG, observamos que a maior perda de informação ocorreu na variável idade (IG= 71,7%). Isto ocorreu devido sua categorização em faixas etárias (intervalos de 10 anos). O impacto desta categorização também se reflete no ENU (69,9%) e no GG (87,9%). As demais variáveis parecem ter sofrido impacto apenas da supressão dos casos, com exceção da cor da pele (ENU=96,1%), em razão do agrupamento em dois grupos: “branca e preta” e “outras”.

Tabela 3 Métricas para medir o efeito dos identificadores indiretos nos conjuntos D e Dz

Métrica	Sexo	Idade	Escolaridade	Cor da Pele	Estado Civil
IG	98,5	71,7	98,5	98,5	98,5
GG	98,5	87,9	98,5	98,5	98,5
ENU	98,6	69,9	98,2	96,1	98,5
Valores Suprimidos	1,4	7,6	1,4	1,4	1,4

IG: Intensidade de Generalização, GG: Granularidade Geral, ENU: Entropia Não Uniforme

Fonte: Elaborado pelos autores, 2022.

Para as métricas de aferição do modelo, apresentamos na Tabela 4 a AUC-ROC, acurácia, sensibilidade e especificidade comparando os resultados da IA aplicada nos conjuntos *D, Di* e *Dz*. De um modo geral parece não haver um grande impacto da anonimização na performance dos modelos. As maiores diferenças ocorreram na acurácia e na sensibilidade com menores valores no conjunto anonimizado.

Tabela 4 - Performance dos modelos treinados com 3 conjuntos de dados

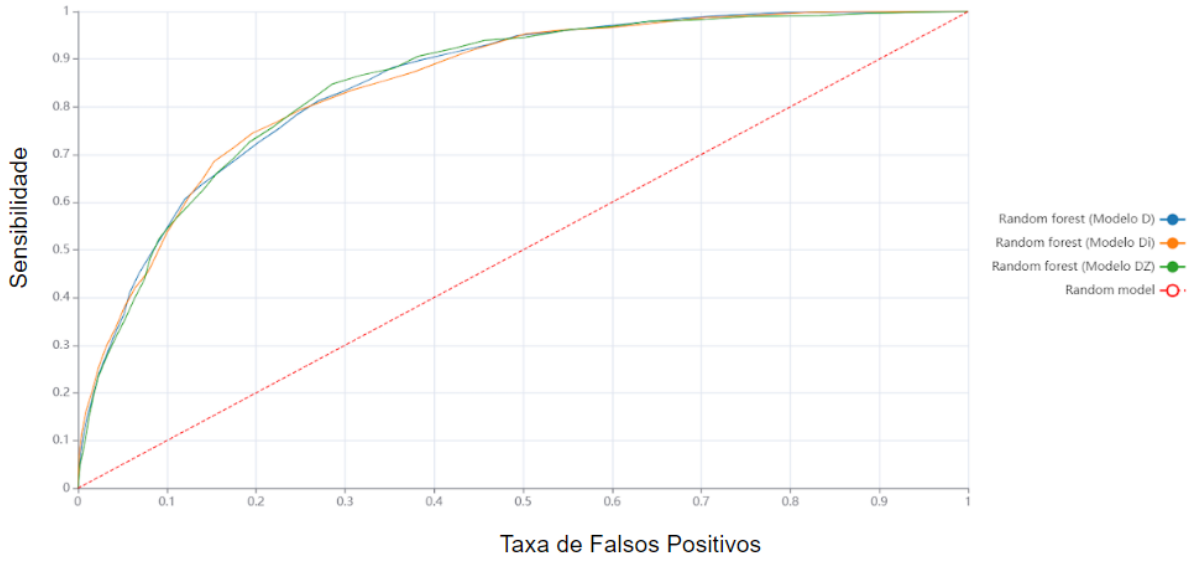
Métrica	Conjunto D	Conjunto Di	Conjunto Dz
AUC-ROC	86,2	85,7	85,5
Acurácia	74,5	72,4	72,9
Sensibilidade	86,1	85,2	83,5
Especificidade	71,3	70,9	70,7

AUC-ROC: *Area Under the Curve* - Receiver Operating Characteristic Curve

Fonte: Elaborado pelos autores, 2022.

A Figura 2 mostra as curvas ROC dos três conjuntos. Nela podemos observar que nenhum conjunto mostra uma curva que seja sempre superior às demais.

Figura 2 - Curva ROC do preditor dos três cenários



Fonte: Elaborado pelos autores, 2022.

DISCUSSÃO

A principal importância deste trabalho é apresentar uma receita (metodologia) para que grupos que trabalham com bases de dados hospitalares possam organizar os seus dados de acordo com a sua necessidade de garantia de privacidade. Além disso, os resultados mostram que dados anonimizados podem ter uma grande redução do risco de identificação (RRId máximo de 100% para 5%) com alguma perda de informação (maior perda: ENU de 98,6% para 69,9%), mas praticamente sem perda de performance (maior perda: sensibilidade de 86,1% para 83,5%). Isso indica que é possível desenvolver novas aplicações da IA a partir de conjuntos anonimizados, garantindo privacidade sem perder qualidade. Outros estudos apresentam resultados divergentes relacionados a isto, LEFEVRE (2008) (LEFEVRE; DEWITT; RAMAKRISHNAN, 2008) mostra menos de 1% de perda na acurácia em modelos treinados utilizando grandes bases de dados reais com e sem o k-anonimato e PURDAM e ELLIOT (2007) apresentaram diferenças superiores a 10% na acurácia dos modelos treinados com e sem o k-anonimato (PURDAM, [*s. d.*]). Entretanto, nosso estudo soma a outras evidências de que a utilidade dos dados dependerá do tipo de uso que está sendo proposto. (DOMINGO-FERRER; REBOLLO-MONEDERO, 2009)

Fizemos este trabalho devido a necessidade de disponibilizar internamente dados para pesquisa. Mas para tornar dados hospitalares abertos na internet, a exemplo do MIMIC, foi feito um grande esforço do MIT em parceria com a Phillips e o hospital Beth-Israel de Boston (MARK, 2016). Implementar em um centro de dados hospitalar esta receita de forma sistematizada para o compartilhamento de dados entre equipes científicas deverá potencializar a reprodutibilidade dos experimentos ao mesmo tempo em que diminuirá os riscos contra a quebra de privacidade durante o desenvolvimento de modelos de IA que utilizam grandes volumes de dados. Uma revisão sistemática

de ataques de reidentificação demonstra que somente quando nenhuma desidentificação foi realizada no dados ou a desidentificação aplicada não era consistente ou baseada nas melhores práticas que os conjuntos foram re-identificados com uma taxa alta de sucesso. No entanto, quando os padrões de reidentificação apropriados são usados, o risco de reidentificação é realmente reduzido e conseguimos reproduzir este efeito durante nosso experimento. Portanto, corroboramos com a evidência que existe hoje e que sugere usar os padrões atuais e as melhores práticas para fornecer alguma proteção contra a reidentificação (EL EMAM; DANKAR, 2008).

Além do uso na pesquisa, sabemos que os hospitais utilizam registros secundários em estudos observacionais para gerar indicadores, modelos estatísticos e auxiliar na tomada de decisão, gerando resultados para o paciente e para a gestão (MAYO CLINIC, [*s. d.*]; LIU *et al.*, 2013; LOWE *et al.*, 2009) Dois grandes exemplos de investimentos em bases de registros secundários vem da Mayo Clinic (rede hospitalar que é referência em informatização na área da saúde nos EUA) e da gigante de tecnologia Apple. A primeira, construiu sua solução de dados denominada *Enterprise Data Trust* em 2010, com o objetivo de consolidar e padronizar semanticamente os dados de diferentes sistemas utilizados em unidades da instituição espalhadas pelos Estados Unidos (CHUTE *et al.*, 2010). A Apple, anunciou em 2019 a inclusão da saúde no centro da sua estratégia de negócios, além de criar sua própria base de dados para pesquisa (HEALTHCARE, [*s. d.*]). Mostramos que é possível propor uma receita genérica para preparação de dados, que simplifica a compreensão dos requisitos necessários para utilizar modelos de IA, também capaz de ser utilizada por outros centros de dados menores que a Apple e a Mayo Clinic. Cria-se uma alternativa para levar os modelos de IA até os hospitais, e não o contrário.

Nosso estudo possui limitações, pois utilizou dados de apenas um hospital, embora de grande porte, e implementamos apenas o modelo de privacidade do k-anonimato para o tratamento de somente 5 atributos considerados identificadores indiretos. Segundo SWEENEY (2002) os dados hospitalares quando são expostos indevidamente, são alvos de ataques do tipo jornalista que invariavelmente tem sucesso na reidentificação de ao menos um registro. Logo, proteger bases de registros hospitalares com o k-anonimato é o ponto de partida para a redução de riscos de reidentificação. Novos estudos são necessários para investigar o efeito dos demais algoritmos e modelos de ataque contra a privacidade e também para o preparo de outros tipos de dados, como: som e imagens médicas, entre outros.

CONCLUSÃO

Seguindo esta receita de preparo, cada pesquisador ou centro de pesquisa pode desde já estabelecer os seus parâmetros de privacidade, informando ao *Data Protection Officer (DPO*) da instituição indicadores mensuráveis encontrados em seus dados e principalmente fornecendo para Agência Nacional de Proteção de Dados (ANPD) um insumo quantitativo sobre estes riscos, para que assim possam ser definidas de forma clara quais são os tipos de riscos e os limites aceitos.

MATERIAL SUPLEMENTAR

ANEXO I - Hiperparâmetros do modelo de aprendizado de máquina para aferição

Algoritmo:	Random forest classification
Critério de Divisão:	Gini
Número de Árvores::	200
Parâmetro bootstrap:	Yes
Profundidade Máxima:	12
Min samples per leaf	7
Min samples to split	21
Método de validação:	Validação cruzada
k-folds:	5
Random seed	1337

Configurações do Computador utilizado no experimento:

Sistema Linux Ubuntu LTS 20.01 64 bits

Intel Core i7 (octa-core) com 16 Gb RAM

Solid-State-Disk com 500 Gb

ANEXO II - Quadro de variáveis do modelo de aferição

Ordem de Importância no Modelo	Variável	Tipo de Atributo	Descrição	Coeficiente de Gini
1	Amplitude de Distribuição dos Glóbulos Vermelhos (Máximo)	Dado Pessoal Sensível	Valor Máximo observado em laboratório da Amplitude de Distribuição dos Glóbulos Vermelhos (RDW do inglês - *Red Cell Distribution Width*).	9,5%
2	Linfócitos (Mínimo)	Dado Pessoal Sensível	Valor Absoluto Mínimo observado em laboratório dos Linfócitos. O valor "absoluto" denota o número total de Linfócitos e não o seu percentual relativo.	7,45%
3	Albumina (Mínimo)	Dado Pessoal Sensível	Valor Mínimo da Albumina. observado em laboratório	5,5%
4	Gravidade	Dado Pessoal Sensível	Escore de Gravidade avaliado na Emergência	5,1%
5	Ureia (Máximo)	Dado Pessoal Sensível	Ureia Valor Máximo observado em Laboratório	5,1%
6	Hemoglobina (Mínimo)	Dado Pessoal Sensível	Hemoglobina Valor Mínimo observado em Laboratório	3,8%
7	Eletrólitos (Contagem)	Dado Pessoal Sensível	Eletrólitos quantidade de exames realizados	3,8%
8	Idade	Identificador Indireto	Idade do paciente na data de internação	3,8%
9	Exames Sangue (Contagem)	Dado Pessoal Sensível	Exames quantidade de exames no sangue realizados	3,2%
10	Eritrócito (Mínimo)	Dado Pessoal Sensível	Eritrócito Valor Mínimo Observado em Laboratório	3,1%
11	Proteína C Reativa (Máximo)	Dado Pessoal Sensível	Proteína C Reativa Valor Máximo Observado em Laboratório	2,8%
12	Frequência Cardíaca (Máximo)	Dado Pessoal Sensível	Frequência Cardíaca Máxima registrada pela Enfermagem	2,6%
13	Sódio (Máximo)	Dado Pessoal Sensível	Sódio Valor Máximo Observado em Laboratório	2,6%
14	Sinais Vitais count	Dado Pessoal Sensível	Quantidade de vezes em que os Sinais Vitais do paciente foram registrados pela enfermagem	2,6%
15	Frequência Cardíaca (Mínimo)	Dado Pessoal Sensível	Frequência cardíaca mínima registrada pela enfermagem	2,4%
16	Frequência Respiratória (Máximo)	Dado Pessoal Sensível	Frequência respiratória máxima registrada pela enfermagem	2,4%
17	Leucócitos (Máximo)	Dado Pessoal Sensível	Leucócitos Valor Máximo Observado em Laboratório	2,2%
18	Plaquetas (Mínimo)	Dado Pessoal Sensível	Plaquetas sanguíneas Valor mínimo Observado em Laboratório	1,9%
19	Tempo de Protrombina (Máximo)	Dado Pessoal Sensível	Tempo de Protrombina em Segundos. Valor Máximo Observado em Laboratório	1,9%
20	Potássio (Máximo)	Dado Pessoal Sensível	Potássio Valor Máximo Observado em Laboratório	1,9%
21	Volume Corpuscular Médio	Dado Pessoal Sensível	Volume Corpuscular Médio	1,9%

	(Máximo)		(VCM), valor Máximo Observado em Laboratório	
22	Hemoglobina Corpuscular Média (Máximo)	Dado Pessoal Sensível	Hemoglobina Corpuscular Média (HCM). Valor Máximo Observado em Laboratório	1,2%
23	Bicarbonato (Mínimo)	Dado Pessoal Sensível	Bicarbonato Valor Mínimo Observado em Laboratório	1,1%
24	SpO2 (Mínimo)	Dado Pessoal Sensível	SpO2 Valor mínimo Observado em Laboratório	1,1%
25	Hemoglobina (Contagem)	Dado Pessoal Sensível	Quantidade de exames de Hemoglobina realizados pelo paciente	1,1%
26	Aspartato (Máximo)	Dado Pessoal Sensível	Aspartato aminotransferase ou transaminase oxalacética (AST ou TGO). Valor Máximo Observado em Laboratório	1,1%
27	Tempo de Tromboplastina Parcial Ativada (Máximo)	Dado Pessoal Sensível	Tempo de Tromboplastina Parcial Ativada (TTPA). Valor Máximo Observado em Laboratório	1%
28	Potássio (Mínimo)	Dado Pessoal Sensível	Potássio Valor mínimo Observado em Laboratório	1%
29	Gasometria Arterial PCO2 (Mínimo)	Dado Pessoal Sensível	Gasometria Arterial PCO2 Valor mínimo da Gasometria Arterial	1%
30	Neutrófilos Bastonete Absoluto (Máximo)	Dado Pessoal Sensível	Neutrófilos Bastonete Absoluto Valor Máximo Observado em Laboratório	0,9%
31	Fosfatase Alcalina (Máximo)	Dado Pessoal Sensível	Fosfatase Alcalina Valor Máximo Observado em Laboratório	0,9%
32	Pressão Arterial Sistólica (Mínimo)	Dado Pessoal Sensível	Pressão Arterial Sistólica mínima registrada pela enfermagem	0,9%
33	Pressão Arterial Diastólica (Mínimo)	Dado Pessoal Sensível	Pressão arterial diastólica mínima registrada pela enfermagem	0,9%
34	Tax (Máximo)	Dado Pessoal Sensível	Temperatura axilar máxima registrada pela enfermagem	0,9%
35	Gasometria Arterial HCO3 (Mínimo)	Dado Pessoal Sensível	Gasometria Arterial HCO3 Valor Máximo da gasometria arterial	0,9%
36	Gasometria Arterial O2SAT (Mínimo)	Dado Pessoal Sensível	Gasometria Arterial O2SAT mínimo da gasometria arterial	0,9%
37	Gasometria Arterial PO2 (Mínimo)	Dado Pessoal Sensível	Gasometria Arterial PO2 mínimo da gasometria arterial	0,9%
38	Alanina (Máximo)	Dado Pessoal Sensível	Alanina Valor Máximo Observado em Laboratório	0,9%
39	Gasometria Arterial PCO2 (Máximo)	Dado Pessoal Sensível	Gasometria Arterial PCO2 Valor Máximo da gasometria arterial	0,9%
40	Gasometria Arterial HCO3 (Máximo)	Dado Pessoal Sensível	Gasometria Arterial HCO3 Valor Máximo da gasometria arterial	0,9%
41	Tomografia (S/N)	Dado Pessoal Sensível	Indica a realização de tomografia computadorizada	0,9%
42	Gasometria Arterial EB (Máximo)	Dado Pessoal Sensível	Gasometria Arterial EB Valor Máximo Gasometria Arterial	0,7%
43	Gasometria Arterial PH (Mínimo)	Dado Pessoal Sensível	Gasometria Arterial PH Valor mínimo da gasometria arterial	0,6%
44	Frequência Respiratória (Mínimo)	Dado Pessoal Sensível	Frequência respiratória mínima registrada pela enfermagem	0,5%

45	Magnésio (Mínimo)	Dado Pessoal Sensível	Magnésio Valor mínimo Observado em Laboratório	0,4%
46	Lactato Desidrogenase (Máximo)	Dado Pessoal Sensível	Lactato Desidrogenase (LDH) Valor Máximo Observado em Laboratório	0,4%
47	Gasometria Venosa SAT (Mínimo)	Dado Pessoal Sensível	Gas_Ven_SAT Valor mínimo de gasometría venosa	0,4%
48	Glicose (Máximo)	Dado Pessoal Sensível	Glicose Valor Máximo Observado em Laboratório	0,3%
49	Gasometria Venosa PO2 (Mínimo)	Dado Pessoal Sensível	Gasometria Venosa PO2 mínima	0,3%
50	Glicose (Mínimo)	Dado Pessoal Sensível	Glicose mínima Observado em Laboratório	0,3%
51	Gasometria Venosa PH (Mínimo)	Dado Pessoal Sensível	Gasometria Venosa PH mínima da gasometria	0,1%
52	Gasometria Venosa CO2T (Máximo)	Dado Pessoal Sensível	Gasometria Venosa CO2T máximo da gasometria	0,1%
53	Procedimento (S/N)	Dado Pessoal Sensível	Indica a realização de procedimento cirúrgico no paciente	0,1%
54	Eletrocardiograma (S/N)	Dado Pessoal Sensível	Indica a realização de eletrocardiograma no paciente	0,1%
55	Gasometria Venosa HCO3 (Máximo)	Dado Pessoal Sensível	Gasometria Venosa HCO3 Máximo da gasometria	0,1%
56	Estado Civil	Identificador Indireto	Estado civil do paciente	0,1%
57	Troponina (Máximo)	Dado Pessoal Sensível	Troponina Valor Máximo Observado em Laboratório	0,1%
58	Sexo	Identificador Indireto	Sexo biológico do paciente	0,1%
59	Tabagista (S/N)	Dado Pessoal Sensível	Indica se o paciente é tabagista ou não	0,1%
60	Gasometria Venosa EB (Máximo)	Dado Pessoal Sensível	Gasometria Venosa EB Valor Máximo	0,1%
61	HbA1c (Máximo)	Dado Pessoal Sensível	HbA1c Valor Máximo Observado em Laboratório	0,1%
63	Escolaridade	Identificador Indireto	Nível de ensino do paciente	0,1%
66	Creatinofosfoquinase (Máximo)	Dado Pessoal Sensível	Creatinofosfoquinase (CPK) Valor Máximo Observado em Laboratório	0,1%
67	Cor da Pele	Identificador Indireto	Cor da pele autodeclarada pelo paciente	0,1%
70	Metamielócitos Absoluto (Máximo)	Dado Pessoal Sensível	Metamielócitos Absoluto Valor Máximo Observado em Laboratório	0,1%
72	Fibrinogênio (Mínimo)	Dado Pessoal Sensível	Fibrinogênio Valor mínimo Observado em Laboratório	0,1%
76	Ddímeros (Máximo)	Dado Pessoal Sensível	Ddímeros Valor Máximo Observado em Laboratório	0,1%
77	Transferrina (Máximo)	Dado Pessoal Sensível	Transferrina Valor Máximo Observado em Laboratório	0,1%

Referências

ADVANCED COHORT EXPLORER DATA RETRIEVAL - CENTER FOR CLINICAL AND TRANSLATIONAL SCIENCE (CCATS) - MAYO CLINIC. [*S. l.*], [*s. d.*]. Disponível em: http://www.mayo.edu/ctsa/resources/consultative-resources/advanced-cohort-explorer-data-retrieval. Acesso em: 22 set. 2017.

ALSHUGRAN, Tariq; DICHTER, Julius. Extracting and modeling the privacy requirements from HIPAA for healthcare applications. *Em*: , 2014. **IEEE Long Island Systems, Applications and Technology (LISAT) Conference 2014**. [*S. l.: s. n.*], 2014. p. 1–5.

BHONGE, Himanshu N.; AMBAT, Monish K.; CHANDAVARKAR, B. R. An Experimental Evaluation of SHA-512 for Different Modes of Operation. *Em*: , 2020. **2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)**. [*S. l.: s. n.*], 2020. p. 1–6.

BOONSTRA, Albert; VERSLUIS, Arie; VOS, Janita F. J. Implementing electronic health records in hospitals: a systematic literature review. **BMC health services research**, [*s. l.*], v. 14, p. 370, 2014.

BRAJER, Nathan *et al.* Prospective and External Evaluation of a Machine Learning Model to Predict In-Hospital Mortality of Adults at Time of Admission. **JAMA Network Open**, [*s. l.*], v. 3, n. 2, p. e1920733–e1920733, 2020. Disponível em: Acesso em: 20 nov. 2020.

CHUTE, Christopher G. *et al.* The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. **Journal of the American Medical Informatics Association: JAMIA**, [*s. l.*], v. 17, n. 2, p. 131–135, 2010.

COMMITTEE ON STRATEGIES FOR RESPONSIBLE SHARING OF CLINICAL TRIAL DATA; BOARD ON HEALTH SCIENCES POLICY; INSTITUTE OF MEDICINE. **Concepts and Methods for De-identifying Clinical Trial Data**. [*S. l.*]: National Academies Press (US), 2015. *E-book*. Disponível em: Acesso em: 21 abr. 2022.

DANKAR, Fida Kamal; EL EMAM, Khaled. A method for evaluating marketer re-identification risk. *Em*: , Article 28., 2010, Lausanne, Switzerland. **Proceedings of the 2010 EDBT/ICDT Workshops**. New York, NY, USA: Association for Computing Machinery, 2010. p. 1–10. Disponível em: Acesso em: 9 maio 2022.

DATAIKU | YOUR PATH TO ENTERPRISE AI. [*S. l.*], [*s. d.*]. Disponível em: https://www.dataiku.com/dss/. Acesso em: 28 mar. 2019.

DOMINGO-FERRER, Josep; REBOLLO-MONEDERO, David. Measuring risk and utility of anonymized data using information theory. *Em*: , 2009, Saint-Petersburg, Russia. **Proceedings of the 2009 EDBT/ICDT Workshops**. New York, NY, USA: Association for Computing Machinery, 2009. p. 126–130. Disponível em: Acesso em: 9 maio 2022.

DWORK, Cynthia. Differential Privacy: A Survey of Results. *Em*: , 2008. **Theory and Applications of Models of Computation**. [*S. l.*]: Springer Berlin Heidelberg, 2008. p. 1–19.

EICHER, Johanna *et al.* A comprehensive tool for creating and evaluating privacy-preserving biomedical prediction models. **BMC medical informatics and decision making**, [*s. l.*], v. 20, n. 1, p. 29, 2020.

EL EMAM, Khaled; DANKAR, Fida Kamal. Protecting privacy using k-anonymity. **Journal of the American Medical Informatics Association: JAMIA**, [*s. l.*], v. 15, n. 5, p. 627–637, 2008.

HCPA. **Sistema AGHUse - Portal Hospital de Clínicas de Porto Alegre**. [*S. l.*], 2014. Disponível em: https://www.hcpa.edu.br/institucional/tecnologia-da-informacao/institucional-sistema-aghuse. Acesso em: 22 fev. 2018.

HEALTHCARE. [*S. l.*], [*s. d.*]. Disponível em: https://www.apple.com/healthcare/. Acesso em: 24 out. 2019.

HRIPCSAK, George *et al.* Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. **Studies in health technology and informatics**, [*s. l.*], v. 216, p. 574–578, 2015.

HYPPÖNEN, Hannele *et al.* Impacts of structuring the electronic health record: a systematic review protocol and results of previous reviews. **International journal of medical informatics**, [*s. l.*], v. 83, n. 3, p. 159–169, 2014.

IYENGAR, Vijay S. Transforming data to satisfy privacy constraints. *Em*: , 2002, Edmonton, Alberta, Canada. **Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining**. New York, NY, USA: Association for Computing Machinery, 2002. p. 279–288. Disponível em: Acesso em: 21 abr. 2022.

KAHN, Michael G. *et al.* A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. **EGEMS (Washington, DC)**, [*s. l.*], v. 4, n. 1, p. 1244, 2016.

KAWAMURA, Takao. Interpretação de um teste sob a visão epidemiológica: eficiência de um teste. **Arquivos brasileiros de cardiologia**, [*s. l.*], v. 79, n. 4, p. 437–441, 2002. Disponível em: Acesso em: 19 out. 2017.

KIMBALL, Ralph; CASERTA, Joe. **The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data**. [*S. l.*]: John Wiley & Sons, 2011.

LEFEVRE, K.; DEWITT, D. J.; RAMAKRISHNAN, R. Mondrian Multidimensional K-Anonymity. *Em*: , 2006. **22nd International Conference on Data Engineering (ICDE'06)**. [*S. l.: s. n.*], 2006. p. 25–25.

LEFEVRE, Kristen; DEWITT, David J.; RAMAKRISHNAN, Raghu. Workload-aware anonymization techniques for large-scale datasets. **ACM Trans. Database Syst.**, New York, NY, USA, v. 33, n. 3, p. 1–47, 2008.

LIU, Hongfang *et al.* An information extraction framework for cohort identification using electronic health records. **AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science**, [*s. l.*], v. 2013, p. 149–153, 2013.

LOWE, Henry J. *et al.* STRIDE--An integrated standards-based translational research informatics platform. **AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium**, [*s. l.*], v. 2009, p. 391–395, 2009.

MACHANAVAJJHALA, Ashwin *et al. L*-diversity: Privacy beyond *k*-anonymity. **ACM transactions on knowledge discovery from data**, New York, NY, USA, v. 1, n. 1, p. 3 – es, 2007.

MARK, Roger. The Story of MIMIC. *Em*: MIT CRITICAL DATA (org.). **Secondary Analysis of Electronic Health Records**. Cham: Springer International Publishing, 2016. p. 43–49.

ORGANIZATION, International Standardization. **ISO 8601: 2004 (E): Data Elements and Interchange Formats, Information Interchange, Representation of Dates and Times**. [*S. l.*]: ISO, 2004.

PANOV, Panče; SOLDATOVA, Larisa N.; DŽEROSKI, Sašo. Generic ontology of datatypes. **Information sciences**, [*s. l.*], v. 329, p. 900–920, 2016.

PEDREGOSA, Fabian *et al.* Scikit-learn: Machine learning in Python. **the Journal of machine Learning research**, [*s. l.*], v. 12, p. 2825–2830, 2011.

PRASSER, Fabian *et al.* ARX--A Comprehensive Tool for Anonymizing Biomedical Data. **AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium**, [*s. l.*], v. 2014, p.

984–993, 2014.

PRASSER, Fabian *et al.* Flexible data anonymization using ARX—Current status and challenges ahead. **Software: practice & experience**, [*s. l.*], v. 50, n. 7, p. 1277–1304, 2020.

PRASSER, Fabian; BILD, Raffael; KUHN, Klaus A. A Generic Method for Assessing the Quality of De-Identified Health Data. **Studies in health technology and informatics**, [*s. l.*], 2016. Disponível em: https://www.researchgate.net/publication/317006066_A_Generic_Method_for_Assessing_the_Quality_of_De-Identified_Health_Data. Acesso em: 19 maio 2022.

PURDAM, M. J. Elliot Dr. TM2 - A Case Study of the Impact of Statistical Disclosure Control on Data Quality in the UK Samples of Anonymised Records. [*s. l.*], Disponível em: https://research.cbs.nl/casc/deliv/5-D2.pdf.

REWATKAR, Liladhar R.; LANJEWAR, Ujwal A. Necessity to design of new DBMS platforms for data analysis in market-oriented cloud computing: properties and limitations of data analysis. **International Journal of Computational Intelligence Research**, [*s. l.*], v. 6, p. 449+, 2010.

SILVA, Helen Ribeiro da. Adoção de tecnologia em hospitais: o caso da adoção do sistema AGHU pelos hospitais universitários do Brasil. [*s. l.*], 2016. Disponível em: http://www.bdm.unb.br/handle/10483/13973.

SHEWHART, W. **Statistical Method from the Viewpoint of Quality Control. By Walter A. Shewhart, edited by W. E. Deming. The Graduate School, The Department of Agriculture, Washington, D. C. 155 pages**. [*S. l.: s. n.*], 1940. Disponível em: http://dx.doi.org/10.1086/286647.

SORENSEN, H. T.; SABROE, S.; OLSEN, J. A framework for evaluation of secondary data sources for epidemiological research. **International journal of epidemiology**, [*s. l.*], v. 25, n. 2, p. 435–442, 1996.

SWEENEY, Latanya. ACHIEVING k-ANONYMITY PRIVACY PROTECTION USING GENERALIZATION AND SUPPRESSION. **International Journal of Uncertainty, Fuzziness and Knowlege-Based Systems**, [*s. l.*], v. 10, n. 05, p. 571–588, 2002.

SZLOSEK, Donald A.; FERRETT, Jonathan. Using machine learning and Natural Language Processing algorithms to automate the evaluation of clinical decision support in electronic medical record systems. **EGEMS (Washington, DC)**, [*s. l.*], v. 4, n. 3, p. 1222, 2016.

VAZ, Tiago Andres. Modelo de dados para treinamento de inteligência artificial na pesquisa em saúde: um estudo prático sobre infecções hospitalares. [*s. l.*], 2017. Disponível em: https://lume.ufrgs.br/handle/10183/181275. Acesso em: 11 nov. 2020.

WANG, S. V. *et al.* **Transparency and Reproducibility of Observational Cohort Studies Using Large Healthcare Databases**. [*S. l.: s. n.*], 2016. Disponível em: http://dx.doi.org/10.1002/cpt.329.

WICKHAM, Hadley. Tidy Data. **Journal of statistical software**, [*s. l.*], v. 59, p. 1–23, 2014. Disponível em: Acesso em: 2 maio 2022.

WICKHAM, Hadley; GROLEMUND, Garrett. **R for Data Science: Import, Tidy, Transform, Visualize, and Model Data**. [*S. l.*]: “O’Reilly Media, Inc.”, 2016.

REFERÊNCIAS

1. ABADI, Martín *et al.* **Deep Learning with Differential Privacy**. 2016. Disponível em: https://doi.org/10.1145/2976749.2978318
2. ABHYANKAR, Swapna; DEMNER-FUSHMAN, Dina; MCDONALD, Clement J. Standardizing clinical laboratory data for secondary use. **Journal of biomedical informatics**, *[S. l.]*, v. 45, n. 4, p. 642–650, 2012.
3. ABOAB, Jerôme *et al.* A “datathon” model to support cross-disciplinary collaboration. **Science translational medicine**, *[S. l.]*, v. 8, n. 333, p. 333ps8–ps333ps8, 2016.
4. ABRAHÃO, Maria Tereza Fernandes; NOBRE, Moacyr Roberto Cuce; MADRIL, Pablo Jorge. O estado da arte em pesquisa observacional de dados de saúde: A iniciativa OHDSI. **Sociedade Brasileira de Computação**, *[S. l.]*, 2019. Disponível em: https://sol.sbc.org.br/livros/index.php/sbc/catalog/book/29/98/248-1. Acesso em: 17 nov. 2020.
5. ACKOFF, Russell L. From data to wisdom. **Journal of Applied Systems Analysis**, *[S. l.]*, v. 16, n. 1, p. 3–9, 1989.
6. **Alistair’s Website!**. . *[s. l.]*, *[s. d.]*. Disponível em: https://alistairewj.github.io/. Acesso em: 11 nov. 2020.
7. ALMEIDA FILHO, Naomar de. Bases históricas da Epidemiologia. **Cadernos de Saúde Pública**, *[S. l.]*, v. 2, n. 3, p. 304–311, 1986.
8. A prize for discoveries past, present and future. **Nature Machine Intelligence**, *[S. l.]*, v. 1, n. 5, p. 201–201, 2019.
9. ARCENIO, Luiz Fernando Stopa. Integração dos Data Warehouse do AGHU das filiais MEC/EBSERH com Pentaho Data Integration. **Anais SULCOMP**, *[S. l.]*, v. 7, 2015. Disponível em: http://periodicos.unesc.net/sulcomp/article/view/1813
10. **Art. 5, inc. X da Constituição Federal de 88**. . *[s. l.]*, *[s. d.]*. Disponível em: https://www.jusbrasil.com.br/topicos/10730704/inciso-x-do-artigo-5-da-constituicao-federal-de-1988. Acesso em: 9 nov. 2020.
11. AZEVEDO, Walter Fernandes de *et al.* Complicações da gravidez na adolescência: revisão sistemática da literatura. **Einstein** , *[S. l.]*, v. 13, n. 4, p. 618–626, 2015.
12. AZUR, Melissa J. *et al.* Multiple imputation by chained equations: what is it and how does it work? **International journal of methods in psychiatric research**, *[S. l.]*, v. 20, n. 1, p. 40–49, 2011.
13. BENNETT, C. J.; RAAB, C. D. The governance of privacy: Policy instruments in global perspective. *[S. l.]*, 2017. Disponível em: https://content.taylorfrancis.com/books/download?dac=C2017-0-57746-0&isbn=9781351775489&format=googlePreviewPdf
14. BENSON, Tim; GRIEVE, Grahame. **Principles of Health Interoperability:**

SNOMED CT, HL7 and FHIR. *[S. l.]*: Springer, Cham, 2016. *E-book*.
15. BITTAR, O. J. N.; BICZYK, M.; SERINOLLI, M. I. Sistemas de informação em saúde e sua complexidade. **de Administração em …**, *[S. l.]*, 2018. Disponível em: http://www.cqh.org.br/ojs-2.4.8/index.php/ras/article/view/77
16. BONAWITZ, Keith *et al.* Practical Secure Aggregation for Privacy-Preserving Machine Learning. *In*: 2017, **Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security**. : ACM, 2017. p. 1175–1191.
17. BOONE, M. Dustin *et al.* The organizational structure of an intensive care unit influences treatment of hypotension among critically ill patients: A retrospective cohort study. **Journal of critical care**, *[S. l.]*, v. 33, p. 14–18, 2016.
18. BOUSSADI, Abdelali; ZAPLETAL, Eric. **A Fast Healthcare Interoperability Resources (FHIR) layer implemented over i2b2**. *[S. l.: s. n.]* Disponível em: https://doi.org/10.1186/s12911-017-0513-6
19. BOWIE, Norman E.; JAMAL, Karim. Privacy Rights on the Internet: Self-Regulation or Government Regulation? **Business ethics quarterly: the journal of the Society for Business Ethics**, *[S. l.]*, v. 16, n. 3, p. 323–342, 2006.
20. BRACCI, F.; CORRADI, A.; FOSCHINI, L. Database security management for healthcare SaaS in the Amazon AWS Cloud. *In*: 2012, **2012 IEEE Symposium on Computers and Communications (ISCC)**. *[S. l.: s. n.]* p. 000812–000819.
21. BRAJER, Nathan *et al.* Prospective and External Evaluation of a Machine Learning Model to Predict In-Hospital Mortality. *[S. l.]*, 2019. Disponível em: https://doi.org/10.1101/19000133
22. BRASIL. **LGPD: A Lei Geral de Proteção de Dados Pessoais**. *[s. l.]*, 2018. Disponível em: http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2018/Lei/L13709.htm. Acesso em: 11 fev. 2019.
23. BUNEMAN, Peter. Semistructured Data. **Information Science**, *[S. l.]*, 1997. Disponível em: http://www.cis.upenn.edu/~db
24. CAMERON, D.; JONES, I. G. John Snow, the broad street pump and modern epidemiology. **International journal of epidemiology**, *[S. l.]*, v. 12, n. 4, p. 393–396, 1983.
25. CHAGAS, Elenita Teresinha Charão *et al.* Planejamento de órteses e próteses e materiais especiais para pacientes SUS no sistema de agendamento cirúrgico. **Clinical and biomedical research. Porto Alegre**, *[S. l.]*, 2017. Disponível em: https://www.lume.ufrgs.br/bitstream/handle/10183/171035/001051073.pdf?sequence=1
26. **Character encodings: Essential concepts**. . *[s. l.]*, *[s. d.]*. Disponível em: https://www.w3.org/International/articles/definitions-characters/. Acesso em: 9 nov. 2020.
27. CHECKLAND, Peter; HOLWELL, Sue. **Information, Systems and Information Systems: Making Sense of the Field**. New York, NY, USA: John Wiley & Sons, Inc., 1998. *E-book*.
28. CHEN, Junqiao *et al.* The validity of synthetic clinical data: a validation study of a leading synthetic data generator (Synthea) using clinical quality measures. **BMC medical informatics and decision making**, *[S. l.]*, v. 19, n. 1, p. 44, 2019.
29. CHEN, Peter Pin-Shan. The entity-relationship model—toward a unified view of data. **ACM Trans. Database Syst.**, New York, NY, USA, v. 1, n. 1, p. 9–36, 1976.
30. CHOI, Edward *et al.* **Doctor AI: Predicting Clinical Events via Recurrent Neural Networks**. 2015. Disponível em: http://arxiv.org/abs/1511.05942v11
31. CHOI, Edward *et al.* GRAM: Graph-based Attention Model for Healthcare Representation Learning. *In*: 2017, New York, NY, USA. **Proceedings of the 23rd**

ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery, 2017. p. 787–795.
32. CHUTE, Christopher G. *et al.* The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. **Journal of the American Medical Informatics Association: JAMIA**, *[S. l.]*, v. 17, n. 2, p. 131–135, 2010.
33. CIRIANI, V. *et al.* Microdata Protection. *In*: YU, Ting; JAJODIA, Sushil (org.). **Secure Data Management in Decentralized Systems**. Boston, MA: Springer US, 2007. p. 291–321. *E-book*.
34. COLLINS, Francis S.; TABAK, Lawrence A. Policy: NIH plans to enhance reproducibility. **Nature**, *[S. l.]*, v. 505, n. 7485, p. 612–613, 2014.
35. CORREIA, Lourani Oliveira dos Santos; PADILHA, Bruna Merten; VASCONCELOS, Sandra Mary Lima. Métodos para avaliar a completitude dos dados dos sistemas de informação em saúde do Brasil: uma revisão sistemática. **Ciência & Saúde Coletiva**, *[S. l.]*, v. 19, n. 11, p. 4467–4478, 2014.
36. DAHMEN, Jessamyn; COOK, Diane. SynSys: A Synthetic Data Generation System for Healthcare Applications. **Sensors** , *[S. l.]*, v. 19, n. 5, 2019. Disponível em: https://doi.org/10.3390/s19051181
37. DALE, Nell; WALKER, Henry M. A Classification of Data Types. **Computer Science Education**, *[S. l.]*, v. 3, n. 3, p. 223–232, 1992.
38. DATA, Mit Critical. **Secondary Analysis of Electronic Health Records**. *[S. l.]*: Springer, 2016. *E-book*.
39. DE BRUIJN, Berry *et al.* Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. **Journal of the American Medical Informatics Association: JAMIA**, *[S. l.]*, v. 18, n. 5, p. 557–562, 2011.
40. **De-Identification Software Package v1.1**. . *[s. l.]*, 2007. Disponível em: https://physionet.org/content/deid/1.1/. Acesso em: 6 dez. 2019.
41. DENLEY, Andrew; FOULSHAM, Mark; HITCHEN, Brian. **GDPR principles**. *[S. l.: s. n.]* Disponível em: https://doi.org/10.4324/9780429449970-3
42. DE STATISTIQUE APPLIQUÉE, Gilbert Saporta Chaire; CEDEX, F. 75141 Paris. Data Mining and Official Statistics. *[S. l.]*, *[s. d.]*. Disponível em: http://cedric.cnam.fr/fichiers/RC184.pdf
43. DORA, Jose Miguel *et al.* Development of a local relative value unit to measure radiologists' computed tomography reporting workload. **Journal of medical imaging and radiation oncology**, *[S. l.]*, v. 60, n. 6, p. 714–719, 2016.
44. DUKE. DukeCathR - Documentation for users. *[S. l.]*, 2016. Disponível em: https://dcri.org/wp-content/uploads/2016/10/DukeCathR_documentation-for-users_161110.pdf
45. **Duke Databank Exhibit**. . *[s. l.]*, *[s. d.]*. Disponível em: http://digitaldukemed.mc.duke.edu/databank/. Acesso em: 13 nov. 2019.
46. DUNN, Olive Jean; CLARK, Virginia A. **Basic Statistics: A Primer for the Biomedical Sciences**. *[S. l.]*: John Wiley & Sons, 2009. *E-book*.
47. DWORK, Cynthia. Differential Privacy. *In*: 2006, Berlin, Heidelberg. **Proceedings of the 33rd International Conference on Automata, Languages and Programming - Volume Part II**. Berlin, Heidelberg: Springer-Verlag, 2006. p. 1–12.
48. DWORK, Cynthia; ROTH, Aaron. The Algorithmic Foundations of Differential Privacy. **Foundations and Trends in Theoretical Computer Science**, *[S. l.]*, v. 9, n. 3–4, p. 211–407, 2014.
49. EL EMAM, Khaled; DANKAR, Fida Kamal. Protecting privacy using k-anonymity. **Journal of the American Medical Informatics Association: JAMIA**, *[S. l.]*, v. 15, n. 5, p. 627–637, 2008.

50. ERLINGSSON, Úlfar; PIHUR, Vasyl; KOROLOVA, Aleksandra. **RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response**. 2014. Disponível em: http://arxiv.org/abs/1407.6981
51. FERGUSON, T. Bruce, Jr *et al.* A decade of change—risk profiles and outcomes for isolated coronary artery bypass grafting procedures, 1990--1999: a report from the STS National Database Committee and the Duke Clinical Research Institute. **The Annals of thoracic surgery**, *[S. l.]*, v. 73, n. 2, p. 480–489, 2002.
52. FLAUSINO, Vinícius Silva; OTHERS. Cultura e poder na organização hospitalar: as relações de poder na implantação da EBSERH em um hospital universitário. *[S. l.]*, 2015. Disponível em: http://repositorio.ufu.br/handle/123456789/12011
53. FLOURIS, Andreas D.; DUFFY, Jack. Applications of artificial intelligence systems in the analysis of epidemiological data. **European journal of epidemiology**, *[S. l.]*, v. 21, n. 3, p. 167–170, 2006.
54. FREEMAN, R. *et al.* Advances in electronic surveillance for healthcare-associated infections in the 21st Century: a systematic review. **The Journal of hospital infection**, *[S. l.]*, v. 84, n. 2, p. 106–119, 2013.
55. FUNG, Benjamin C. M. *et al.* **Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques**. 1st. ed. *[S. l.]*: Chapman & Hall/CRC, 2010. *E-book*.
56. GARZA, Maryam *et al.* Evaluating common data models for use with a longitudinal community registry. **Journal of biomedical informatics**, *[S. l.]*, v. 64, p. 333–341, 2016.
57. GENRO, Bruna Pasqualini *et al.* 25 anos de Bioética Clínica no HCPA: Um pioneirismo que se renova. **Clinical & Biomedical Research**, *[S. l.]*, v. 38, n. 3, *[s. d.]*. Disponível em: https://www.seer.ufrgs.br/hcpa/article/view/87256
58. GIONIS, A.; TASSA, T. k-Anonymization with Minimal Loss of Information. **IEEE transactions on knowledge and data engineering**, *[S. l.]*, v. 21, n. 2, p. 206–219, 2009.
59. GOLLE, Philippe. Revisiting the Uniqueness of Simple Demographics in the US Population. *In*: 2006, New York, NY, USA. **Proceedings of the 5th ACM Workshop on Privacy in Electronic Society**. New York, NY, USA: ACM, 2006. p. 77–80.
60. GONCALVES, Andre *et al.* Generation and evaluation of synthetic patient data. **BMC medical research methodology**, *[S. l.]*, v. 20, n. 1, p. 108, 2020.
61. GRAMMATICOS, Philip C.; DIAMANTIS, Aristidis. Useful known and unknown views of the father of modern medicine, Hippocrates and his teacher Democritus. **Hellenic journal of nuclear medicine**, *[S. l.]*, v. 11, n. 1, p. 2–4, 2008.
62. GREENBERG, Andy *et al.* Apple's "Differential Privacy" Is About Collecting Your Data—But Not Your Data. **Wired**, *[S. l.]*, 2016. Disponível em: https://www.wired.com/2016/06/apples-differential-privacy-collecting-data/. Acesso em: 17 dez. 2019.
63. GRUBB, Alex F. *et al.* Tobacco smoking in patients with heart failure and coronary artery disease: A 20-year experience at Duke University Medical Center. **American heart journal**, *[S. l.]*, v. 230, p. 25–34, 2020.
64. GUTIERREZ, Marco Antonio. **Experiência do InCor em Sistemas de Informações em Saúde**. *[s. d.]*. - Instituto do Coração, *[s. l.]*, *[s. d.]*. Disponível em: http://www.telessaude.uerj.br/resource/goldbook/pdf/26.pdf
65. GYMREK, Melissa *et al.* Identifying personal genomes by surname inference. **Science**, *[S. l.]*, v. 339, n. 6117, p. 321–324, 2013.
66. **HAI Data and Statistics | HAI | CDC**. . *[s. l.]*, *[s. d.]*. Disponível em: https://www.cdc.gov/hai/surveillance/index.html. Acesso em: 14 out. 2017.
67. HALLIDAY, S. Death and miasma in Victorian London: an obstinate belief. **BMJ** , *[S. l.]*, v. 323, n. 7327, p. 1469–1471, 2001.

68. HARRIS, P. J. *et al.* Survival in medically treated coronary artery disease. **Circulation**, *[S. l.]*, v. 60, n. 6, p. 1259–1269, 1979.
69. HCA-FAB. **Implantação do Sistema AGHUse no HCA - Força Aérea Brasileira**. *[s. l.]*, 2018. Disponível em: http://www2.fab.mil.br/hca/index.php/2014-12-11-17-51-57/301-implantacao-do-sistema-aghuse-no-hca. Acesso em: 11 fev. 2019.
70. HCPA. **Sistema AGHUse - Portal Hospital de Clínicas de Porto Alegre**. *[s. l.]*, 2014. Disponível em: https://www.hcpa.edu.br/institucional/tecnologia-da-informacao/institucional-sistema-aghuse. Acesso em: 22 fev. 2018.
71. HENDERSON, D. A. Epidemiology in the global eradication of smallpox. **International journal of epidemiology**, *[S. l.]*, v. 1, n. 1, p. 25–30, 1972.
72. HOCHMAN, Bernardo *et al.* [Research designs]. **Acta cirurgica brasileira / Sociedade Brasileira para Desenvolvimento Pesquisa em Cirurgia**, *[S. l.]*, v. 20 Suppl 2, p. 2–9, 2005.
73. HOLOHAN, Naoise *et al.* **(k,ε)-Anonymity: k-Anonymity with ε-Differential Privacy**. 2017. Disponível em: http://arxiv.org/abs/1710.01615
74. HOMER, Nils *et al.* Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. **PLoS genetics**, *[S. l.]*, v. 4, n. 8, p. e1000167, 2008.
75. HRIPCSAK, George *et al.* Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. **Studies in health technology and informatics**, *[S. l.]*, v. 216, p. 574–578, 2015.
76. HULLEY, Stephen B. *et al.* Delineando a pesquisa clínica: uma abordagem epidemiológica. *In*: **Delineando a pesquisa clínica: uma abordagem epidemiológica**. *[S. l.]*: Artmed, 2008. *E-book*.
77. HUMPHEYS, N. A. Vital Statistics: a Memorial Volume of Selections from the Reports and Writings of Willian Farr MD, DCL, CB, FRS. **London: Edward Stanford**, *[S. l.]*, p. 250–330, 1885.
78. ICHIHARA, Maria Yuri Travassos; BARRETO, Maurício Lima; OTHERS. V Encontros Pré-ConfOA: dados abertos, ciência de dados aplicada à saúde, CIDACS: potencialidades e desafios do Centro de Integração de Dados e Conhecimentos para a Saúde-CIDACS. *[S. l.]*, 2017. Disponível em: https://www.arca.fiocruz.br/handle/icict/22726
79. IENCA, Marcello *et al.* Considerations for ethics review of big data health research: A scoping review. **PloS one**, *[S. l.]*, v. 13, n. 10, p. e0204937, 2018.
80. INSTITUTE FOR HEALTH METRICS AND EVALUATION (IHME). **Findings from the Global Burden of Disease Study 2017**. *[S. l.]*: Seattle, WA: IHME, 2018. Disponível em: http://www.healthdata.org/sites/default/files/files/policy_report/2019/GBD_2017_Booklet.pdf.
81. JENAL, Sabine; ÉVORA, Yolanda Dora Martinez. Revisão de literatura: Implantação de Prontuário Eletrônico do Paciente. **Journal of health informatics in developing countries**, *[S. l.]*, v. 4, n. 4, 2012. Disponível em: http://www.jhi-sbis.saude.ws/ojs-jhi/index.php/jhi-sbis/article/view/216. Acesso em: 23 out. 2019.
82. JOHNSON, Alistair E. W. *et al.* Machine Learning and Decision Support in Critical Care. **Proceedings of the IEEE. Institute of Electrical and Electronics Engineers**, *[S. l.]*, v. 104, n. 2, p. 444–466, 2016 a.
83. JOHNSON, Alistair E. W. *et al.* MIMIC-III, a freely accessible critical care database. **Scientific data**, *[S. l.]*, v. 3, p. 160035, 2016 b.

84. JOHNSON, Alistair Ew *et al.* The MIMIC Code Repository: enabling reproducibility in critical care research. **Journal of the American Medical Informatics Association: JAMIA**, *[S. l.]*, v. 25, n. 1, p. 32–39, 2018.
85. JOHNSON, Alistair E. W. *et al.* **MIMIC-CXR: A large publicly available database of labeled chest radiographs**. 2019. Disponível em: http://arxiv.org/abs/1901.07042
86. KEET, C. Maria *et al.* The Data Mining OPtimization Ontology. **Journal of Web Semantics**, *[S. l.]*, v. 32, p. 43–53, 2015.
87. KELLEY, Al; POHL, Ira. **A Book on C; Programming in C**. 3rd. ed. Redwood City, CA, USA: Benjamin-Cummings Publishing Co., Inc., 1994. *E-book*.
88. KHOURY, Muin J.; IOANNIDIS, John P. A. Medicine. Big data meets public health. **Science**, *[S. l.]*, v. 346, n. 6213, p. 1054–1055, 2014.
89. KIMBALL, Ralph *et al.* **The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses**. *[S. l.]*: John Wiley & Sons, 1998. *E-book*.
90. KLEISIARIS, Christos F.; SFAKIANAKIS, Chrisanthos; PAPATHANASIOU, Ioanna V. Health care practices in ancient Greece: The Hippocratic ideal. **Journal of medical ethics and history of medicine**, *[S. l.]*, v. 7, p. 6, 2014.
91. KLÖSGEN, W. Types and forms of data. **Handbook of Data Mining and Knowledge Discovery**, *[S. l.]*, 2002. Disponível em: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.39.6211&rep=rep1&type=pdf
92. KOHLMAYER, Florian; PRASSER, Fabian; KUHN, Klaus A. The cost of quality: Implementing generalization and suppression for anonymizing biomedical data with minimal information loss. **Journal of biomedical informatics**, *[S. l.]*, v. 58, p. 37–48, 2015.
93. KRIEGER, Nancy. **Epidemiology and the People's Health: Theory and Context**. *[S. l.]*: OUP USA, 2011. *E-book*.
94. KUNTZ, Richard E. *et al.* Individual Patient-Level Data Sharing for Continuous Learning: A Strategy for Trial Data Sharing. **NAM Perspectives**, *[S. l.]*, 2019. Disponível em: https://pdfs.semanticscholar.org/f07e/9f6314c8bed61af77825f561604f7f4dc997.pdf
95. KUSHIDA, Clete A. *et al.* Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. **Medical care**, *[S. l.]*, v. 50 Suppl, p. S82–S101, 2012.
96. LAVATER, Johann Caspar. **Physiognomische Fragmente zur Beförderung der Menschenkenntnis und Menschenliebe. [Faksimiledruck nach der Ausg. 1775-1778]**. *[S. l.]*: Orell Füssli, 1969. *E-book*.
97. LAVIGNE, Maxime *et al.* A population health perspective on artificial intelligence. **Healthcare management forum / Canadian College of Health Service Executives = Forum gestion des soins de sante / College canadien des directeurs de services de sante**, *[S. l.]*, v. 32, n. 4, p. 173–177, 2019.
98. LECUN, Yann; BENGIO, Yoshua; HINTON, Geoffrey. Deep learning. **Nature**, *[S. l.]*, v. 521, n. 7553, p. 436–444, 2015.
99. LEFEVRE, Kristen; DEWITT, David J.; RAMAKRISHNAN, Raghu. Workload-aware anonymization techniques for large-scale datasets. **ACM Trans. Database Syst.**, New York, NY, USA, v. 33, n. 3, p. 1–47, 2008.
100. LENZ, Sylvia Tamara. Michael J. Campbell, David Machin and Stephen J. Walters (2007): Medical Statistics, a Textbook for the Health Sciences, 4th edition. **Statistical Papers**, *[S. l.]*, v. 50, n. 1, p. 217–218, 2009.
101. LIMA, Daniel M. *et al.* Transforming Two Decades of ePR Data to OMOP CDM for Clinical Research. **Studies in health technology and informatics**, *[S. l.]*, v. 264,

p. 233–237, 2019.
102. LIU, Yu; WANG, Ting; FENG, Jianhua. A Semantic Information Loss Metric for Privacy Preserving Publication. *In*: 2010, **Database Systems for Advanced Applications**. : Springer Berlin Heidelberg, 2010. p. 138–152.
103. LOWE, Henry J. *et al.* STRIDE--An integrated standards-based translational research informatics platform. **AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium**, *[S. l.]*, v. 2009, p. 391–395, 2009.
104. LUO, Yuan *et al.* Segment convolutional neural networks (Seg-CNNs) for classifying relations in clinical notes. **Journal of the American Medical Informatics Association: JAMIA**, *[S. l.]*, v. 25, n. 1, p. 93–98, 2018.
105. MACIEL, Daiane Aparecida; FERREIRA, Deborah Pimenta; DE FÁTIMA MARIN, Heimar. A utilização de terminologias para representar os procedimentos e intervenções. **Journal of health informatics in developing countries**, *[S. l.]*, v. 11, n. 4, 2019. Disponível em: http://www.jhi-sbis.saude.ws/ojs-jhi/index.php/jhi-sbis/article/view/671. Acesso em: 10 nov. 2020.
106. MA, Fenglong *et al.* KAME: Knowledge-based Attention Model for Diagnosis Prediction in Healthcare. *In*: 2018, New York, NY, USA. **Proceedings of the 27th ACM International Conference on Information and Knowledge Management**. New York, NY, USA: Association for Computing Machinery, 2018. p. 743–752.
107. MALIN, Bradley; BENITEZ, Kathleen; MASYS, Daniel. Never too old for anonymity: a statistical standard for demographic data sharing via the HIPAA Privacy Rule. **Journal of the American Medical Informatics Association: JAMIA**, *[S. l.]*, v. 18, n. 1, p. 3–10, 2011.
108. MARGÓCSY, Dániel; SOMOS, Mark; JOFFE, Stephen N. Vesalius annotations and the rise of early modern medicine. **The Lancet**, *[S. l.]*, v. 393, n. 10173, p. 738–739, 2019.
109. MARINHO, Fatima *et al.* Burden of disease in Brazil, 1990–2016: a systematic subnational analysis for the Global Burden of Disease Study 2016. **The Lancet**, *[S. l.]*, v. 392, n. 10149, p. 760–775, 2018.
110. MARK, Roger. The Story of MIMIC. *In*: MIT CRITICAL DATA (org.). **Secondary Analysis of Electronic Health Records**. Cham: Springer International Publishing, 2016. p. 43–49. *E-book*.
111. MARTÍNEZ LLUÍS, Sergio; OTHERS. **Ontology based semantic anonymisation of microdata**. *[s. d.]*. - Universitat Rovira i Virgili, *[s. l.]*, *[s. d.]*. Disponível em: https://www.tdx.cat/handle/10803/108961
112. MATEO-SANZ, Josep Maria; SEBÉ, Francesc; DOMINGO-FERRER, Josep. Outlier Protection in Continuous Microdata Masking. *In*: 2004, **Privacy in Statistical Databases**. : Springer Berlin Heidelberg, 2004. p. 201–215.
113. **Measuring Utility and Information Loss — SDC Practice Guide documentation**. . *[s. l.]*, *[s. d.]*. Disponível em: https://sdcpractice.readthedocs.io/en/latest/utility.html. Acesso em: 4 dez. 2019.
114. MENTZ, Robert J. *et al.* Heart failure with preserved ejection fraction: comparison of patients with and without angina pectoris (from the Duke Databank for Cardiovascular Disease). **Journal of the American College of Cardiology**, *[S. l.]*, v. 63, n. 3, p. 251–258, 2014.
115. MIOTTO, Riccardo *et al.* Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. **Scientific reports**, *[S. l.]*, v. 6, n. 1, p. 26094, 2016.
116. MIT CRITICAL DATA. **Secondary Analysis of Electronic Health Records**. *[S. l.]*: Springer, Cham, 2016. *E-book*.

117. MOHASSEL, P.; ZHANG, Y. SecureML: A System for Scalable Privacy-Preserving Machine Learning. *In*: 2017, **2017 IEEE Symposium on Security and Privacy (SP)**. *[S. l.: s. n.]* p. 19–38.
118. NOY, Natalya F.; MCGUINNESS, Deborah L.; OTHERS. **Ontology development 101: A guide to creating your first ontology**. *[S. l.]*: Stanford knowledge systems laboratory technical report KSL-01-05 and …, 2001. Disponível em: https://protege.stanford.edu/publications/ontology_development/ontology101.pdf
119. OECD; OECD. **OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data (Summary in Portuguese)**. *[S. l.: s. n.]* Disponível em: https://doi.org/10.1787/9789264196391-sum-pt
120. **OHDSI Studies**. . *[s. l.]*, *[s. d.]*. Disponível em: https://data.ohdsi.org/OhdsiStudies/. Acesso em: 13 nov. 2020.
121. OLIVEIRA, Diego Farias. **A IMPLANTAÇÃO DE UM SISTEMA DE GESTÃO DE CUSTOS NO HOSPITAL UNIVERSITÁRIO PELA EBSERH: UM ESTUDO DE CASO COM UTILIZAÇÃO DO PMBOK**. *[S. l.: s. n.]* Disponível em: https://doi.org/10.21450/rahis.v13i3.3172
122. **OMOP-CDM Conversion and Anonymization of National Health Insurance Service-National Sample Cohort**. . *[s. l.]*, *[s. d.]*. Disponível em: https://www.ohdsi.org/2019-us-symposium-showcase-17/. Acesso em: 3 nov. 2020.
123. **Open Data Portal (ODP)**. . *[s. l.]*, *[s. d.]*. Disponível em: https://www.nhsbsa.nhs.uk/open-data-portal-odp. Acesso em: 13 nov. 2020.
124. **Open Health Data**. . *[s. l.]*, *[s. d.]*. Disponível em: https://openhealthdata.metajnl.com/. Acesso em: 26 dez. 2019.
125. PANETH, Nigel. Assessing the contributions of John Snow to epidemiology: 150 years after removal of the broad street pump handle. **Epidemiology** , *[S. l.]*, v. 15, n. 5, p. 514–516, 2004.
126. PANOV, Panče; SOLDATOVA, Larisa N.; DŽEROSKI, Sašo. Generic ontology of datatypes. **Information sciences**, *[S. l.]*, v. 329, p. 900–920, 2016.
127. PARKER, Richard B. A definition of privacy. *In*: **Privacy**. *[S. l.]*: Routledge, 2017. p. 83–104. *E-book*.
128. PATRÍCIO, Camila Mendes *et al.* O prontuário eletrônico do paciente no sistema de saúde brasileiro: uma realidade para os médicos? **Scientia medica**, *[S. l.]*, v. 21, n. 3, 2011. Disponível em: http://revistaseletronicas.pucrs.br/ojs/index.php/scientiamedica/article/viewFile/8723/6722&g
129. PEDREGOSA, Fabian *et al.* Scikit-learn: Machine learning in Python. **the Journal of machine Learning research**, *[S. l.]*, v. 12, p. 2825–2830, 2011.
130. PIRES, Fábio Antero. **Ambiente para extração de informação epidemiológica a partir da mineração de dez anos de dados do Sistema Público de Saúde**. 2011. - Universidade de São Paulo, *[s. l.]*, 2011. Disponível em: http://www.teses.usp.br/teses/disponiveis/5/5131/tde-08122011-145701/en.php
131. PIZA, Felipe Maia de Toledo *et al.* Assessing team effectiveness and affective learning in a datathon. **International journal of medical informatics**, *[S. l.]*, v. 112, p. 40–44, 2018.
132. POLLARD, Tom *et al.* **Open Data in Health Care**. *[S. l.: s. n.]* Disponível em: https://doi.org/10.5334/ban.h
133. POLLARD, Tom J. *et al.* tableone: An open source Python package for producing summary statistics for research papers. **JAMIA Open**, *[S. l.]*, v. 1, n. 1, p. 26–31, 2018.
134. POPKIN, Gabriel. Data sharing and how it can benefit your scientific career. **Nature**, *[S. l.]*, v. 569, n. 7756, p. 445–447, 2019.

135. **Portal CMD 1.0**. . *[s. l.]*, *[s. d.]*. Disponível em: https://conjuntominimo.saude.gov.br/#/cmd. Acesso em: 13 nov. 2019.
136. POWELL, John; BUCHAN, Iain. **Electronic health records should support clinical research**. **Journal of medical Internet research**, 2005.
137. PRASSER, Fabian; KOHLMAYER, Florian. Putting Statistical Disclosure Control into Practice: The ARX Data Anonymization Tool. *In*: GKOULALAS-DIVANIS, Aris; LOUKIDES, Grigorios (org.). **Medical Data Privacy Handbook**. Cham: Springer International Publishing, 2015. p. 111–148. *E-book*.
138. PREVEDELLO, Luciano M. *et al.* Automated Critical Test Findings Identification and Online Notification System Using Artificial Intelligence in Imaging. **Radiology**, *[S. l.]*, v. 285, n. 3, p. 923–931, 2017.
139. **Programa que vai integrar dados de usuários do SUS em todo o país é lançado em Alagoas**. . *[s. l.]*, 2019. Disponível em: https://g1.globo.com/al/alagoas/noticia/2019/11/11/programa-que-vai-integrar-dados-de-usuarios-do-sus-em-todo-o-pais-e-lancado-em-alagoas.ghtml. Acesso em: 13 nov. 2019.
140. QUEIROZ, Maria J.; LINO, Natasha C. Q.; GUSTAVO H M. Uma Ontologia de Domínio para Preservação de Privacidade em Dados Publicados pelo Governo Brasileiro. *In*: 2016, **Anais do XII Simpósio Brasileiro de Sistemas de Informação**. : SBC, 2016. p. 009–016.
141. RAJENDRAN, Keerthana; JAYABALAN, Manoj; RANA, Muhammad Ehsan. A Study on k-anonymity, l-diversity, and t-closeness Techniques focusing Medical Data. *[S. l.]*, v. 17, n. 12, 2017. Disponível em: http://dx.doi.org/. Acesso em: 6 dez. 2019.
142. RAJKOMAR, Alvin *et al.* Scalable and accurate deep learning with electronic health records. **npj Digital Medicine**, *[S. l.]*, v. 1, n. 1, p. 18, 2018.
143. RATWANI, Raj M. *et al.* A usability and safety analysis of electronic health records: a multi-center study. **Journal of the American Medical Informatics Association: JAMIA**, *[S. l.]*, v. 25, n. 9, p. 1197–1201, 2018.
144. RELVA, Dervaneide Santos. **Análise do grau de aceitação do aplicativo AGHU durante sua implantação no Hospital Universitário Onofre Lopes**. 2016. - Universidade Federal do Rio Grande do Norte, *[s. l.]*, 2016. Disponível em: https://monografias.ufrn.br/jspui/handle/123456789/4093
145. RENJIFO, Carlos A. **Exploration, processing and visualization of physiological signals from the ICU**. 2005. - Massachusetts Institute of Technology, *[s. l.]*, 2005. Disponível em: https://dspace.mit.edu/handle/1721.1/33350?show=full. Acesso em: 25 out. 2019.
146. REPS, Jenna M. *et al.* Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. **Journal of the American Medical Informatics Association: JAMIA**, *[S. l.]*, v. 25, n. 8, p. 969–975, 2018.
147. RILEY, James C. **Rising Life Expectancy: A Global History**. *[S. l.]*: Cambridge University Press, 2001. *E-book*.
148. ROCHER, Luc; HENDRICKX, Julien M.; DE MONTJOYE, Yves-Alexandre. Estimating the success of re-identifications in incomplete datasets using generative models. **Nature communications**, *[S. l.]*, v. 10, n. 1, p. 3069, 2019.
149. ROWLEY, Jennifer. The wisdom hierarchy: representations of the DIKW hierarchy. **Journal of Information Science and Engineering**, *[S. l.]*, v. 33, n. 2, p. 163–180, 2007.
150. SAEED, M. *et al.* MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring. **Computers in cardiology**, *[S. l.]*, v. 29, p. 641–644, 2002.

151. SANTOS, Ricardo S. *et al.* Big Data Analytics in a Public General Hospital. *In*: 2016, **Machine Learning, Optimization, and Big Data**. : Springer International Publishing, 2016. p. 433–441.
152. SERPA NETO, Ary *et al.* First Brazilian datathon in critical care. **Revista Brasileira de terapia intensiva**, *[S. l.]*, v. 30, n. 1, p. 6–8, 2018 a.
153. SERPA NETO, Ary *et al.* Mechanical power of ventilation is associated with mortality in critically ill patients: an analysis of patients in two observational cohorts. **Intensive care medicine**, *[S. l.]*, v. 44, n. 11, p. 1914–1922, 2018 b.
154. SHABAN-NEJAD, Arash; MICHALOWSKI, Martin; BUCKERIDGE, David L. Health intelligence: how artificial intelligence transforms population and personalized health. **NPJ digital medicine**, *[S. l.]*, v. 1, p. 53, 2018.
155. SHEET, Issue. Pseudonymisation and anonymisation of data policy. *[S. l.]*, *[s. d.]*. Disponível em: https://www.nhsbsa.nhs.uk/sites/default/files/2017-05/anonymisation-of-data-policy.pdf
156. SHICKEL, B. *et al.* Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. **IEEE Journal of Biomedical and Health Informatics**, *[S. l.]*, v. 22, n. 5, p. 1589–1604, 2018.
157. SHI, Leming *et al.* The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. **Nature biotechnology**, *[S. l.]*, v. 28, n. 8, p. 827–838, 2010.
158. SHORTLIFFE, Edward Hance; PERREAULT, Leslie E. **Medical Informatics: Computer Applications in Health Care**. *[S. l.]*: Addison-Wesley Publishing Company, 1990. *E-book*.
159. SILVA, Helen Ribeiro da. Adoção de tecnologia em hospitais: o caso da adoção do sistema AGHU pelos hospitais universitários do Brasil. *[S. l.]*, 2016. Disponível em: http://www.bdm.unb.br/handle/10483/13973
160. SILVA, João Vitor Ferreira da; OTHERS. A utilização do Sistema e-SUS Hospitalar na farmácia do Hospital Federal de Bonsucesso. *[S. l.]*, 2016. Disponível em: http://app.uff.br/riuff/handle/1/1902
161. SILVA, Valter Ferreira da. Registros eletrônicos de saúde e pesquisa clínica : ferramentas para permitir um uso adequado e descentralizado de informações. *[S. l.]*, 2019. Disponível em: https://lume.ufrgs.br/handle/10183/197764. Acesso em: 11 nov. 2020.
162. SILVA, Pollianna Marys de Souza e.; DE SOUZA E SILVA, Pollianna Marys; DE AUTRAN, Marynice Medeiros Matos. **REPOSITÓRIO DATASUS: ORGANIZAÇÃO E RELEVÂNCIA DOS DADOS ABERTOS EM SAÚDE PARA A VIGILÂNCIA EPIDEMIOLÓGICA**. *[S. l.: s. n.]* Disponível em: https://doi.org/10.21721/p2p.2019v6n1.p50-59
163. SIRAISI, Nancy G. **Medieval and Early Renaissance Medicine: An Introduction to Knowledge and Practice**. *[S. l.]*: University of Chicago Press, 2009. *E-book*.
164. SNEDDON, Tam P.; LI, Peter; EDMUNDS, Scott C. GigaDB: announcing the GigaScience database. **GigaScience**, *[S. l.]*, v. 1, n. 1, p. 11, 2012.
165. SNELL, Elizabeth. **Ensuring Security, Access to Protected Health Information (PHI)**. *[s. l.]*, 2017. Disponível em: https://healthitsecurity.com/features/ensuring-security-access-to-protected-health-information-phi. Acesso em: 29 jul. 2019.
166. **SOAR DATA™ - DCRI**. . *[s. l.]*, *[s. d.]*. Disponível em: https://dcri.org/our-work/analytics-and-data-science/data-sharing/soar-data/. Acesso em: 29 jul. 2019.

167. SOLOVE, Daniel J. **Understanding privacy**. *[S. l.]*: Harvard university press Cambridge, MA, 2008. v. 173*E-book*.
168. SOUZA, Aline Da Cruz Rodrigues *et al.* **Interoperabilidade Técnica Entre Sistemas De Registro Eletrônico Em Saúde Em Organizações Publicas De Saúde Brasileiras**. *[S. l.: s. n.]* Disponível em: https://doi.org/10.5335/rbca.v11i2.8651
169. SOUZA, Antonio Artur de *et al.* Avaliação de Sistemas de Informação: Um Estudo em Organizações Hospitalares. **SOCIEDADE, CONTABILIDADE E GESTÃO**, *[S. l.]*, v. 7, n. 1, 2013. Disponível em: http://atena.org.br/revista/ojs-2.2.3-08/index.php/ufrj/article/viewArticle/1472. Acesso em: 4 nov. 2019.
170. SOUZA, Thais Teles *et al.* Morbidade e mortalidade relacionadas a medicamentos no Brasil: revisão sistemática de estudos observacionais. **Biotecnologia aplicada: revista de la Sociedad Iberolatinoamericana para Investigaciones sobre Interferon y Biotecnologia en Salud**, *[S. l.]*, v. 35, n. 4, 2015. Disponível em: http://serv-bib.fcfar.unesp.br/seer/index.php/Cien_Farm/article/viewArticle/2971. Acesso em: 24 out. 2019.
171. SULLIVAN, June M. **HIPAA: A Practical Guide to the Privacy and Security of Health Data**. *[S. l.]*: American Bar Association, 2004. *E-book*.
172. SUTHERLAND, Ian. John Graunt: A Tercentenary Tribute. **Journal of the Royal Statistical Society. Series A** , *[S. l.]*, v. 126, n. 4, p. 537, 1963.
173. SWEENEY, L. Weaving technology and policy together to maintain confidentiality. **The Journal of law, medicine & ethics: a journal of the American Society of Law, Medicine & Ethics**, *[S. l.]*, v. 25, n. 2-3, p. 98–110, 82, 1997.
174. SWEENEY, Latanya. k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY. **International Journal of Uncertainty, Fuzziness and Knowlege-Based Systems**, *[S. l.]*, v. 10, n. 05, p. 557–570, 2002 a.
175. SWEENEY, Latanya. ACHIEVING k-ANONYMITY PRIVACY PROTECTION USING GENERALIZATION AND SUPPRESSION. **International Journal of Uncertainty, Fuzziness and Knowlege-Based Systems**, *[S. l.]*, v. 10, n. 05, p. 571–588, 2002 b.
176. SWEENEY, Latanya. Only you, your doctor, and many others may know. **Technology Science**, *[S. l.]*, v. 2015092903, n. 9, p. 29, 2015.
177. SWEENEY, Latanya; CROSAS, Mercè; BAR-SINAI, Michael. Sharing Sensitive Data with Confidence: The Datatags System. *[S. l.]*, *[s. d.]*. Disponível em: https://scholar.harvard.edu/files/mercecrosas/files/techsci-datatags-sweeneycrosasbarsinai.pdf
178. SZARVAS, György; FARKAS, Richárd; BUSA-FEKETE, Róbert. State-of-the-art anonymization of medical records using an iterative machine learning framework. **Journal of the American Medical Informatics Association: JAMIA**, *[S. l.]*, v. 14, n. 5, p. 574–580, 2007.
179. TANNA, Sunil. **Binary, Octal and Hexadecimal for Programming & Computer Science (English Edition)**. 1. ed. *[S. l.]*: Answers 2000 Limited, 1 julho 2018. *E-book*.
180. TASNEEM, Asba *et al.* Developing a framework for a comprehensive data sharing program. *In*: 2017, **AMIA**. *[S. l.: s. n.]* Disponível em: https://dukeinformatics.org/wp-content/uploads/2014/10/asba_tasneem_soar_poster_amia2017.pdf
181. THE LABORATORY FOR COMPUTATIONAL PHYSIOLOGY, MIT. **MIMIC Critical Care Database**. *[s. l.]*, *[s. d.]*. Disponível em: https://mimic.physionet.org/.

Acesso em: 29 jul. 2019.
182. THOMAS, Kevin L. *et al.* Racial differences in long-term survival among patients with coronary artery disease. **American heart journal**, *[S. l.]*, v. 160, n. 4, p. 744–751, 2010.
183. TIDKE, Bharat; MEHTA, Rupa; DHANANI, Jenish. A Comprehensive Survey and Open Challenges of Mining Bigdata. *In*: SATAPATHY, Suresh Chandra; JOSHI, Amit (org.). **Information and Communication Technology for Intelligent Systems (ICTIS 2017) - Volume 1**. Cham: Springer International Publishing, 2018. (Smart Innovation, Systems and Technologies).v. 83p. 441–448. *E-book*.
184. TROVATI, Marcello *et al.* **Big-Data Analytics and Cloud Computing: Theory, Algorithms and Applications**. *[S. l.]*: Springer, 2016. *E-book*.
185. TURER, Aslan T. *et al.* Influence of body mass index on the efficacy of revascularization in patients with coronary artery disease. **The Journal of thoracic and cardiovascular surgery**, *[S. l.]*, v. 137, n. 6, p. 1468–1474, 2009.
186. UNICAMP. **AGHUse está próximo de 750 mil acessos em 15 meses | Hospital de Clínicas - UNICAMP**. *[s. l.]*, 2018. Disponível em: https://www.hc.unicamp.br/node/1148. Acesso em: 2018.
187. VAZ, Tiago Andres. Modelo de dados para treinamento de inteligência artificial na pesquisa em saúde: um estudo prático sobre infecções hospitalares. *[S. l.]*, 2017. Disponível em: https://lume.ufrgs.br/handle/10183/181275. Acesso em: 11 nov. 2020.
188. WANG, S. V. *et al.* **Transparency and Reproducibility of Observational Cohort Studies Using Large Healthcare Databases**. *[S. l.: s. n.]* Disponível em: https://doi.org/10.1002/cpt.329
189. WANG, Taowei David *et al.* Extracting insights from electronic health records: case studies, a visual analytics process model, and design recommendations. **Journal of medical systems**, *[S. l.]*, v. 35, n. 5, p. 1135–1152, 2011.
190. WANG, Xiao *et al.* Heterogeneous Graph Attention Network. *In*: 2019, New York, NY, USA. **The World Wide Web Conference**. New York, NY, USA: Association for Computing Machinery, 2019. p. 2022–2032.
191. **What Is Privacy?**. . *[s. l.]*, *[s. d.]*. Disponível em: https://privacyinternational.org/explainer/56/what-privacy. Acesso em: 18 nov. 2019.
192. WIENS, Jenna *et al.* Potential Adverse Effects of Broad-Spectrum Antimicrobial Exposure in the Intensive Care Unit. **Open forum infectious diseases**, *[S. l.]*, v. 5, n. 2, p. ofx270, 2018.
193. WILLENBORG, Leon; DE WAAL, Ton. **Elements of Statistical Disclosure Control**. *[S. l.]*: Springer Science & Business Media, 2012. *E-book*.
194. XU, K. *et al.* Privacy-Preserving Machine Learning Algorithms for Big Data Systems. *In*: 2015, **2015 IEEE 35th International Conference on Distributed Computing Systems**. *[S. l.: s. n.]* p. 318–327.
195. YE, Huimin; CHEN, Elizabeth S. Attribute Utility Motivated k-anonymization of datasets to support the heterogeneous needs of biomedical researchers. **AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium** , *[S. l.]*, v. 2011, p. 1573–1582, 2011.
196. YUN, Seongjun *et al.* Graph Transformer Networks. *In*: (H. Wallach et al., Org.) 2019, **Advances in Neural Information Processing Systems**. : Curran Associates, Inc., 2019. p. 11983–11993.
197. ZHANG, Chuxu *et al.* Heterogeneous Graph Neural Network. *In*: 2019, New York, NY, USA. **Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining**. New York, NY, USA: Association for Computing Machinery, 2019. p. 793–803.
198. ZUCK, D. **John Snow and anaesthesia**. *[S. l.: s. n.]* Disponível em:

https://doi.org/10.1258/jrsm.97.3.153-a