



**UNIVERSIDADE ESTADUAL DA PARAÍBA
CAMPUS I
PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS E TECNOLOGIA EM SAÚDE
MESTRADO EM CIÊNCIAS E TECNOLOGIA EM SAÚDE**

LÍVIA CAROLINE TOMAZ SANTOS

**QUALIDADE DOS DADOS NA INDÚSTRIA DA SAÚDE: PERSPECTIVAS DE BIG
DATA E APRENDIZADO DE MÁQUINA - UMA REVISÃO DE ESCOPO**

**CAMPINA GRANDE - PB
2024**

LÍVIA CAROLINE TOMAZ SANTOS

**QUALIDADE DOS DADOS NA INDÚSTRIA DA SAÚDE: PERSPECTIVAS DE BIG
DATA E APRENDIZADO DE MÁQUINA - UMA REVISÃO DE ESCOPO**

Dissertação apresentada ao Programa de Pós-Graduação em Ciência e Tecnologia em Saúde da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título em Mestre em Ciência e Tecnologia em Saúde.

Área de concentração: Inteligência Artificial aplicada à saúde.

Orientador: Prof. Dr. Frederico Moreira Bublitz

**CAMPINA GRANDE - PB
2024**

É expressamente proibido a comercialização deste documento, tanto na forma impressa como eletrônica. Sua reprodução total ou parcial é permitida exclusivamente para fins acadêmicos e científicos, desde que na reprodução figure a identificação do autor, título, instituição e ano do trabalho.

S237q Santos, Livia Caroline Tomaz.

Qualidade de dados na indústria da saúde [manuscrito] : perspectivas de Big Data e aprendizado de máquina - uma revisão de escopo / Livia Caroline Tomaz Santos. - 2025.

67 p. : il. colorido.

Digitado. Dissertação (Mestrado Profissional em Ciência e Tecnologia em Saúde) - Universidade Estadual da Paraíba, Pró-Reitoria de Pós-Graduação e Pesquisa, 2025. "Orientação : Prof. Dr. Frederico Moreira Bublitz, Departamento de Computação - CCT. "

1. Big Data. 2. Machine Learning. 3. Qualidade de dados.
4. Inteligência artificial aplicada à saúde. I. Título

21. ed. CDD 005.7

LÍVIA CAROLINE TOMAZ SANTOS

QUALIDADE DOS DADOS NA INDÚSTRIA DA SAÚDE: PERSPECTIVAS DE BIG DATA E APRENDIZADO DE MÁQUINA - UMA REVISÃO DE ESCOPO

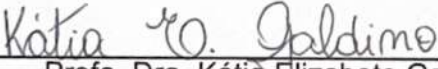
Dissertação apresentada ao Programa de Pós-Graduação em Ciência e Tecnologia em Saúde da Universidade Estadual da Paraíba como requisito para obtenção do título de Mestre em Ciência e Tecnologia em Saúde.

Dissertação aprovada em: 12/12/2024

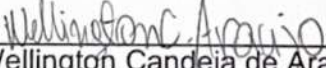
BANCA EXAMINADORA:



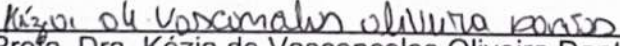
Prof. Dr. Frederico Moreira Bublitz
Universidade Estadual da Paraíba (UEPB)



Profa. Dra. Kátia Elizabete Galdino
Universidade Estadual da Paraíba (UEPB)



Prof. Dr. Wellington Candeia de Araújo
Universidade Estadual da Paraíba (UEPB)



Profa. Dra. Kézia de Vasconcelos Oliveira Dantas
Universidade Estadual da Paraíba (UEPB)

A Deus e aos meus pais, pelo apoio, incentivo
e amor, DEDICO.

AGRADECIMENTOS

Agradeço primeiramente a Deus, por me dar forças a cada dia para continuar e me possibilitar esta conquista.

Aos meus pais, Antonio Carlos e Socorro Palitó, por todo apoio, amor incondicional e dedicação em todos os momentos da minha vida. Agradeço especialmente à minha mãe por sentar ao meu lado e oferecer uma luz nos momentos de bloqueio.

Aos avôs e avós, *in memoriam*, que, embora ausentes fisicamente, continuam sendo uma fonte de inspiração e valores que carrego comigo.

Aos meus irmãos, Carlos Augusto e Bruna Carla, por serem meus melhores amigos ao longo da vida, compartilhando alegrias e desafios.

Ao meu cunhado, Carlos Alejandro, por ser como um irmão e amigo, sempre disposto a ajudar.

Aos meus sobrinhos, Alejandro e Bianca, por proporcionar tanta alegria a cada dia.

Aos meus amigos, Bruno e Vitor, pelos incentivos e momentos de descontração.

Ao Núcleo de Tecnologias Estratégicas em Saúde (NUTES), por todo apoio desde o período de graduação até este momento.

Aos professores do mestrado por compartilhar todo ensinamento, em especial a Profa. Dra. Kátia Galdino e Prof. Dr. Wellington Candeia pelos ótimos conselhos durante a qualificação.

Ao meu orientador, Prof. Dr. Frederico Moreira Bublitz, por toda paciência, acolhimento e guía durante todo o caminho. Muito obrigada!

A todos que contribuíram de forma direta ou indireta para a conclusão desta pesquisa.

RESUMO

Introdução: Tecnologias como *Big Data*, *Machine Learning* e *Deep Learning* possuem grande potencial na área da saúde e podem ser utilizadas para aprimorar o processo de diagnóstico, tratamento e gestão da saúde. Entretanto, para o uso eficaz destas tecnologias é necessário garantir a qualidade dos dados para torná-los capazes de atingir o objetivo proposto. **Objetivo:** Neste contexto, foi realizada uma revisão de escopo para identificar e mapear métodos e *frameworks* utilizados no tratamento e avaliação da qualidade dos dados em aplicações médicas e de saúde. **Métodos e Análise:** A revisão de escopo foi conduzida com base no processo de seis etapas descrito na estrutura proposta por Levac *et al.* em "*Scoping Studies: Advancing the methodology*" e relatado seguindo o *checklist* PRISMA-ScR (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews*). Na busca dos artigos foram utilizados termos associados a *Data Quality*, *Big Data*, *Machine Learning*, *Deep Learning* e *Health* na pesquisa por estudos primários nos bancos de dados Scopus Document Search, IEEE Xplore Digital Library e ACM Digital Library. **Resultados:** Na pesquisa foram incluídos 93 artigos que permitiram responder às questões de pesquisa, na qual forneceram informações sobre os benefícios da aplicação destas tecnologias na indústria da saúde, quais os métodos de pré-processamento dos dados mais utilizados, os seus benefícios e problemas mitigados. Também foram identificadas que as Dimensões de Qualidade de Dados mais abordadas nas pesquisas são confiabilidade, usabilidade e disponibilidade. **Resultados:** O estudo permite entender quais as principais métricas utilizadas para avaliar a qualidade dos dados. E evidenciaram lacunas de conhecimento na relação entre os métodos e as dimensões, dando oportunidade para o desenvolvimento de um framework que possa orientar a seleção e aplicação de métodos adequados para cada dimensão da qualidade.

Palavras-Chave: big data; machine learning; qualidade de dados; saúde.

ABSTRACT

Introduction: Technologies such as Big Data, Machine Learning, and Deep Learning have great potential in the health industry and can be used to improve the process of diagnosis, treatment, and health management. However, to effectively use these technologies, it is necessary to ensure the quality of the data to make them capable of achieving the proposed objective. **Objective:** In this context, a scoping review was carried out to identify and map methods and frameworks used to treat and evaluate data quality in medical and health applications. **Methods and Analysis:** The scoping review was conducted based on the six-step process described in the framework proposed by Levac et al. in "Scoping Studies: Advancing the methodology" and reported following the PRISMA-ScR checklist (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews). In the search for articles, terms associated with data quality, big data, machine learning, deep learning, and health were used to search for primary studies in the Scopus Document Search, IEEE Xplore Digital Library, and ACM Digital Library databases. **Results:** The research included 93 articles that allowed us to answer the research questions, providing information on the benefits of applying these technologies in the health industry, the most commonly used data preprocessing methods, their benefits, and mitigated problems. It was also identified that the Data Quality Dimensions most addressed in the research are reliability, usability, and availability. **Conclusions:** The study lets us understand the main metrics used to assess data quality. It also highlighted gaps in the relationship between methods and dimensions, providing an opportunity to develop a framework to guide the selection and application of appropriate methods for each quality dimension.

Keywords: big data; machine learning; data quality; health.

LISTA DE ILUSTRAÇÕES

Figura 1 - 5 V's Big Data.....	15
Figura 2 - Classificação conceitual dos desafios de Big Data.....	16
Figura 3 - Diferença entre Machine Learning e Deep Learning.....	18
Figura 4 - Um padrão universal de qualidade de big data de duas camadas para avaliação...	22
Figura 5 - Fluxograma PRISMA das etapas de seleção dos artigos.....	28

LISTA DE TABELAS

Tabela 1 - Estudos incluídos na revisão de escopo.....	29
Tabela 2 - Análise dos métodos de qualidade de dados, benefícios e problemas mitigados..	40
Tabela 3 - Análise das Dimensões de Qualidade de Dados.....	43

LISTA DE GRÁFICOS

Gráfico 1 - Quantitativo de estudos por Dimensão de Qualidade de Dados.....	44
---	----

SUMÁRIO

1	INTRODUÇÃO.....	11
2	OBJETIVO.....	14
2.1	Objetivo geral.....	14
2.2	Objetivos específicos.....	14
3	REVISÃO DE LITERATURA.....	15
3.1	Big Data.....	15
3.1.1	<i>Big Data na Saúde.....</i>	17
3.2	Machine Learning e Deep Learning.....	17
3.2.1	<i>Machine Learning e Deep Learning na Saúde.....</i>	19
3.3	Qualidade de Dados.....	20
3.3.1	<i>Dimensões de qualidade de dados.....</i>	21
4	METODOLOGIA.....	23
4.1	Revisão de Escopo.....	23
4.1.1	<i>Identificação da questão de pesquisa.....</i>	24
4.1.2	<i>Identificação dos estudos relevantes.....</i>	25
4.1.3	<i>Seleção dos estudos</i>	25
4.1.4	<i>Mapeamento dos dados.....</i>	26
4.1.5	<i>Agrupar, resumir e relatar os resultados.....</i>	27
5	RESULTADOS.....	28
6	DISCUSSÃO	46
7	CONSIDERAÇÕES FINAIS.....	47
	REFERÊNCIAS.....	48
	APÊNDICE A - PROTOCOLO PUBLICADO.....	59
	ANEXO A - PRISMA ScR.....	66

1 INTRODUÇÃO

Na atualidade, tem-se observado como a tecnologia tem impactado o setor da saúde, auxiliando o trabalho de profissionais e beneficiando pacientes, desde a prevenção, diagnóstico, tratamento e assistência. O crescimento exponencial na quantidade de dados gerados na área da saúde pode ser atribuído a diversos fatores, desde a digitalização dos registros médicos até os avanços tecnológicos como o uso de dispositivos vestíveis, computação em nuvem e internet das coisas (IoT), “o que forçou especialistas em computação a projetar estratégias inovadoras para analisar e interpretar uma quantidade enorme de dados em um determinado período de tempo” (Dash *et al.*, 2019, p.23, tradução nossa).

Com o avanço da era digital, houve o aumento na quantidade de dados obtidos, que, segundo Taylor (2024) em artigo publicado no site Statista, a previsão é que, em 2028 o volume de dados/informações criados, capturados, copiados e consumidos em todo o mundo seja superior a 394 zettabytes. Como forma de obter valor e processar grandes quantidades de dados, faz-se necessário o uso de tecnologias como *Big Data* e *Machine Learning*. *Big Data* refere-se a grandes volumes de conjuntos de dados complexos e diversos (Taylor-Sakyi, 2016). De acordo com Taurion (2013), além de uma grande quantidade de dados, o *Big Data* também é caracterizado por dados diversos e não estruturados que devem ser validados e tratados para que possam ter valor de negócio.

Para Gu *et al.* (2017), tecnologias avançadas como a IoT e redes de sensores permitiram avanços de *Big Data* na Medicina. Algumas das inúmeras aplicações no setor de saúde permitem avanços significativos no diagnóstico, tratamento e gerenciamento de saúde. *Machine Learning* (ML), ou Aprendizado de Máquina, é definido como um modelo computacional capaz de aprender a partir de um conjunto de dados de treinamento, identificando padrões e melhorando sua capacidade de interpretar novos dados para resolver determinados problemas. *Deep Learning* (DL), uma subárea do ML, utiliza redes neurais artificiais inspiradas no cérebro humano para aprender com grandes volumes de dados. DL tem sido fundamental em avanços como reconhecimento de imagem, processamento de linguagem natural e diagnósticos médicos. “Muitos dos avanços atuais do aprendizado de máquina na área da saúde visam apoiar a capacidade do médico ou especialista de fornecer um tratamento mais eficaz aos pacientes com maior qualidade, velocidade e precisão” (Habehh; Gohel, 2021, p.296, tradução nossa).

A relação entre *Big Data* e *Machine Learning* é mutuamente benéfica. Estas tecnologias podem promover e aprimorar uma a outra de forma significativa. *Big Data* fornece o grande volume de dados necessários para treinar modelos de ML, enquanto os algoritmos de ML transformam esses dados em resultados valiosos. O *Deep Learning*, utilizado na análise de dados de *Big Data*, permite encontrar padrões úteis que antes eram desconhecidos (Gheisari; Wang; Bhuiyan, 2017). Embora existam inúmeros benefícios obtidos com o uso destas tecnologias, há também preocupações já que dados de saúde são complexos, variados e muitas vezes inconsistentes, o que pode comprometer a precisão e confiabilidade das análises e modelos de *Machine Learning*. Além disso, no *Big Data*, com o volume considerável de dados e a rapidez do processamento, especialmente com dados não estruturados, que apresentam variações substanciais e são propensos a imprecisões frequentes (Raghupathi; Raghupathi, 2014), podem causar erros nos sistemas de tomada de decisão ou situações de vida ou morte.

A qualidade dos dados em *Big Data* é fundamental, já que dados precisos, consistentes e relevantes permitem que os algoritmos de ML e DL identifiquem padrões corretos e façam previsões mais confiáveis. Em contrapartida, dados de baixa qualidade podem introduzir vieses, gerar resultados enganosos e comprometer a tomada de decisões, especialmente em cenários críticos, como na saúde. “A qualidade da decisão é determinada pela qualidade dos dados, que se refere ao grau de usabilidade dos dados.” (Wang *et al.*, 2024, p.1159, tradução nossa). Garantir a qualidade dos dados não apenas melhora o desempenho dos modelos, mas também fortalece a confiança nas análises e decisões automatizadas, potencializando o valor agregado das tecnologias de *Big Data*.

Cai e Zhu (2015) em “*The Challenges of Data Quality and Data Quality Assessment in the Big Data Era*” identificam cinco Dimensões de Qualidade de Dados (DQD) no contexto de *Big Data*: disponibilidade, usabilidade, confiabilidade, relevância e qualidade da apresentação. A disponibilidade refere-se à facilidade de acesso aos dados quando necessário, enquanto a usabilidade envolve a facilidade com que os dados podem ser aplicados para seus propósitos. A confiabilidade mede o grau de confiança nos dados, considerando sua origem e consistência. A relevância avalia o quão pertinentes os dados são para o problema em questão, e a qualidade da apresentação diz respeito à clareza e organização com que os dados são exibidos, facilitando a interpretação e a tomada de decisões.

Este trabalho de revisão foi baseado no protocolo intitulado “*Data quality and Big Data in the health industry: a scoping review protocol*” (Apêndice A), que serviu como etapa preliminar para a dissertação, ao explorar métodos e estratégias para melhorar a qualidade dos

dados de saúde no contexto de *Big Data*. O escopo foi ampliado para incluir tecnologias como *Machine Learning* e *Deep Learning*, devido à relevância dessas técnicas na análise de grandes volumes de dados de saúde.

Esta revisão de escopo tem como objetivo mapear os métodos utilizados para melhorar a qualidade dos dados de saúde em aplicações que fazem uso de *Big Data* associadas a ML ou DL, apontar os principais benefícios da aplicação destas tecnologias na área da saúde e identificar as DQD apresentadas por Cai e Zhu (2015) tratadas nos artigos inseridos na revisão.

Durante as sequências de buscas para definição dos termos utilizados na *string* de busca, não foram identificadas revisões de escopo cuja ênfase seja compreender e elencar os principais modelos, *frameworks* ou critérios utilizados na qualidade de dados no uso destas tecnologias, cujo foco são os sistemas de saúde. A partir do objetivo proposto, identificamos métodos que podem agregar mais qualidade e segurança ao uso destas tecnologias na indústria da saúde.

2 OBJETIVO

2.1 Objetivo geral

Realizar uma revisão de escopo para identificar os métodos e frameworks aplicados à melhoria da qualidade de dados de saúde em aplicações que utilizam *Big Data*, associados a *Machine Learning* ou *Deep Learning*.

2.2 Objetivos específicos

- Identificar os benefícios da aplicação de *Big Data*, *Machine Learning* e *Deep Learning* na indústria da saúde;
- Buscar os métodos utilizados para a qualidade de dados e os benefícios do uso;
- Identificar quais problemas associados à qualidade dos dados que os métodos solucionam;
- Verificar quais as principais dimensões de qualidade de dados de *Big Data* de acordo com as dimensões apresentadas por Cai e Zhu (2015) são tratadas de forma direta ou indireta para melhorar a qualidade dos dados.

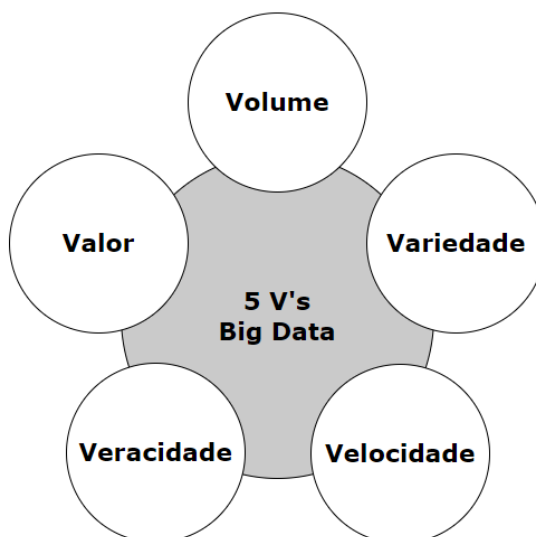
3 REVISÃO DE LITERATURA

3.1 Big Data

Big Data é definido por um grande volume de dados estruturados e não estruturados que são gerados rapidamente, e “que não podem ser processados usando tecnologias tradicionais, métodos de processamento, algoritmos ou quaisquer soluções comerciais prontas para uso” (Krishnan, 2013, p.5, tradução nossa). Ainda de acordo com Krishnan (2013), a popularidade do uso de *Big Data* deve-se às plataformas tecnológicas que surgiram com a capacidade de processar esses diferentes dados. Além da alta capacidade de captura e armazenamento dessa grande quantidade de dados, com o uso do *Big Data* é possível gerenciar e analisar de forma a obter valor destas informações.

O conceito de *Big Data* era caracterizado por 3 V's, volume, velocidade e variedade, porém, com os avanços nos estudos da tecnologia, evoluíram para 5 V's (Alotaibi; Abdullah, 2016), descritos na Figura 1, e definidos por Kumar (2015) como:

Figura 1 - 5 V's Big Data



Fonte: Elaborada pelo autor (2024).

Volume: refere-se a enorme quantidade de dados gerados e coletados.

Variedade: refere-se a diversidade de tipos e formatos de dados gerados.

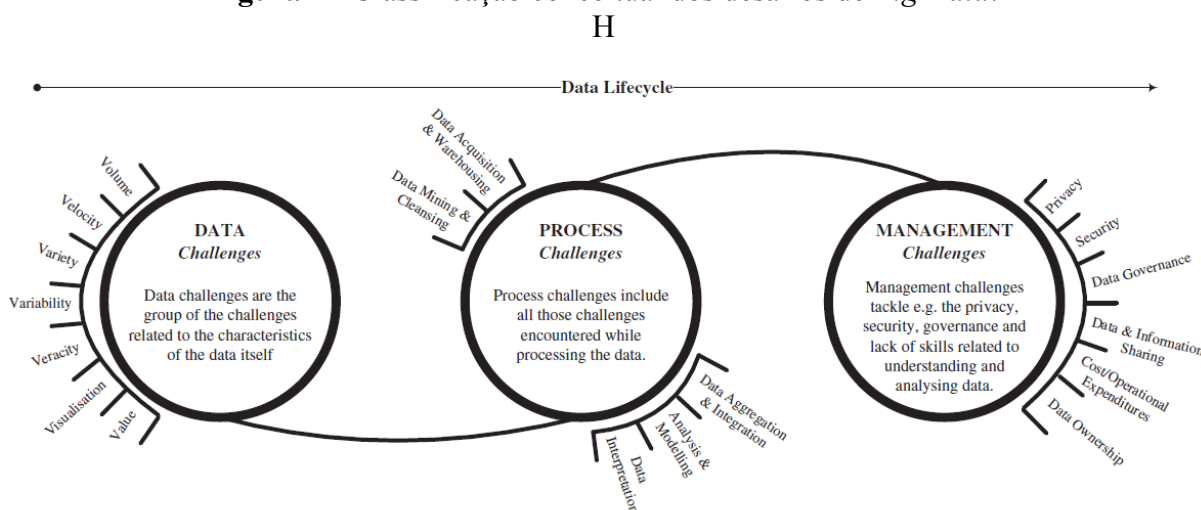
Velocidade: refere-se a velocidade com que os dados são gerados, coletados, armazenados e analisados.

Veracidade: refere-se a confiabilidade e qualidade dos dados.

Valor: refere-se ao valor potencial que pode ser extraído dos dados.

O uso do *Big Data* traz consigo diversos benefícios, como na tomada de decisões, aumento de produtividade, desenvolvimento científico e melhoria da qualidade de vida em diversos setores, desde negócios, saúde, educação, entre outros. Apesar dos benefícios que esta tecnologia proporciona, Sivarajah *et al.* (2017) apresentam os desafios associados ao *Big Data* divididos em três categorias, como apresentado na Figura 2:

Figura 2 - Classificação conceitual dos desafios de *Big Data*.



Fonte: Sivarajah *et al.* (2017).

Desafios de Dados: estão relacionados às próprias características dos dados, apontados nos 5 V's (volume, variedade, velocidade, veracidade e valor). À medida que os estudos sobre o tema avançam, novas características são identificadas e com isso novos desafios.

Desafios de Processo: estão relacionados aos métodos e técnicas necessárias para processar e analisar os dados de forma a obter o resultado esperado. Esses desafios são maiores devido a grande quantidade de dados, diversidade e formatos.

Desafios de Gestão: estão relacionados a aspectos como privacidade, segurança, governança e compartilhamento dos dados.

3.1.1 *Big Data na Saúde*

Big Data tem a capacidade de impactar quase todos os aspectos da sociedade (Gupta; Rani, 2018). Os benefícios do uso do *Big Data* na área da saúde são abrangentes, incluindo melhorias nos serviços de teleassistência, logística, prevenção de doenças e melhoria no atendimento ao paciente. Além de fornecer dados para promover insights valiosos e informações acionáveis a partir de diversas fontes de dados que prevê tendências futuras (Inastrilla, 2022).

Uma das aplicações desta tecnologia foi durante a pandemia de COVID-19, no qual *Big Data* foi utilizado para gerenciar a epidemia global. Haafza *et al.* (2021) identificaram em uma revisão sistemática a aplicação da tecnologia melhorando o processo de diagnóstico, estimativa, prevenção, tomada de decisão e aplicação de medicamentos. Carra *et al.* (2020) apresentam outra aplicação bem-sucedida em modelos prognósticos, preditivos e de classificação baseados em dados obtidos na coleta e análise de dados de UTI.

A utilização de *Big Data* na área da saúde pode resultar em aprimoramentos nos resultados para os pacientes e em cuidados mais eficazes, mas enfrenta desafios como privacidade, qualidade dos dados, inconsistência e instabilidade dos dados. O volume considerável de dados e a rapidez do seu processamento, especialmente com dados não estruturados, apresentam variações substanciais e são propensos a imprecisões frequentes (Luo *et al.*, 2016), causando erros nos sistemas de tomada de decisão ou situações de vida ou morte.

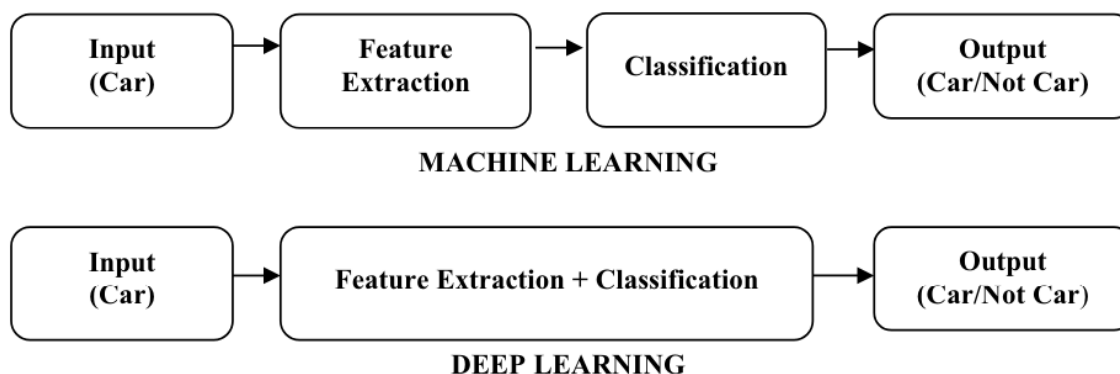
3.2 Machine Learning e Deep Learning

Machine Learning é uma subárea da inteligência artificial, que foca no desenvolvimento de algoritmos e modelos do qual computadores são capazes de aprender através dos dados, permitindo-lhes reconhecer padrões, fazer previsões e tomar decisões com base nos dados obtidos. A precisão do resultado destes modelos tem relação direta ao volume de dados de treinamento usados no seu desenvolvimento (Dargan, 2023).

Deep Learning é caracterizado por redes neurais de múltiplas camadas, o que permite que essas redes aprendam automaticamente padrões complexos e relações não lineares nos dados. As “técnicas convencionais de aprendizado de máquina eram limitadas em sua

capacidade de processar dados naturais em sua forma bruta” (Lecun; Bengio; Hinton, 2015, p. 436, tradução nossa), o DL pode lidar diretamente com dados brutos e extrair características relevantes de maneira eficiente, como observado na Figura 3:

Figura 3 - Diferença entre *Machine Learning* e *Deep Learning*.



Fonte: Dargan *et al.* (2020).

Tanto no *Machine Learning* como *Deep Learning*, dependendo do tipo de problema, da natureza dos dados disponíveis e do objetivo, pode-se escolher um dos três tipos de aprendizado ou, em alguns casos, combinar esses tipos de aprendizado para resolver problemas complexos. Os tipos de aprendizado são:

Aprendizado supervisionado: este tipo é apropriado quando se tem um conjunto de dados de entrada com suas saídas correspondentes. É útil quando se deseja prever ou classificar algo com base em dados históricos.

Aprendizado não supervisionado: abordagem utilizada quando se tem um conjunto de dados não rotulados e se deseja encontrar padrões ou estruturas intrínsecas nos dados.

Aprendizado por reforço: este tipo é utilizado quando o problema envolve tomar sequências de ações na qual visa ensinar um agente a tomar decisões autônomas em ambientes complexos, com base na experiência adquirida através da interação com o ambiente, como em jogos, robótica ou otimização de recursos.

Apesar do potencial transformador do ML e DL, essas tecnologias enfrentam desafios que podem dificultar a sua implementação e eficácia. Alguns dos principais desafios associados ao Machine Learning apresentados por Nithya *et al.* (2023) são:

Quantidade e qualidade dos dados: o desempenho do modelo depende dos dados, já que este pode ser prejudicado em caso de dados imprecisos ou tendenciosos.

Engenharia e seleção de recursos: o desempenho do modelo depende da seleção dos recursos ou variáveis inseridas.

Sobreajuste e subajuste: o modelo se ajusta demais aos dados de treinamento ou é muito simples, resultando em um desempenho insatisfatório com novos dados.

Complexidade e interpretabilidade do modelo: modelos complexos geralmente oferecem melhor previsibilidade, mas podem ser difíceis de interpretar, dificultando entender como as previsões são feitas.

Poder computacional e memória: são recursos computacionais necessários para treinar modelos grandes ou lidar com grandes conjuntos de dados.

Escolha do algoritmo: escolher o melhor algoritmo para um determinado problema, tipo de dados, dificuldade da tarefa e os recursos disponíveis.

Aprendizado contínuo: os modelos devem se adaptar e obter conhecimento em ambientes dinâmicos.

Escalabilidade e implantação: infraestrutura e latência devem ser levadas em consideração para acomodar grandes quantidades de dados ou previsões em tempo real.

Assim como no *Machine Learning*, o *Deep Learning* é influenciado pela disponibilidade e qualidade dos dados utilizados no treinamento dos modelos, mas que é difícil estabelecer a quantidade ideal de dados de treinamento, pois depende tanto da qualidade dos dados quanto da complexidade da tarefa em questão (Choi *et al.*, 2020).

3.2.1 *Machine Learning e Deep Learning na Saúde*

A aplicação de *Machine Learning* e *Deep Learning* na área da saúde tem revolucionado a forma como doenças são diagnosticadas, tratadas e gerenciadas. Rojas *et al.* (2019), apresentam contribuições do uso de ML no setor da saúde, incluindo medicina preventiva, previsão de epidemias e o diagnóstico e cuidado de doenças crônicas como diabetes, câncer e Alzheimer.

Thieme, Belgrave e Doherty (2020) realizaram uma revisão sistemática identificando a aplicação do ML na detecção, diagnóstico e tratamento de problemas associados à saúde mental. Com as técnicas de ML, foi possível avaliar sintomas, acompanhar o avanço da doença e otimizar tratamentos.

Através da aplicação de redes neurais, como Redes Neurais Convolucionais (CNN), que são projetadas especificamente para processamento de imagens, o DL tem sido amplamente utilizado para resolver problemas associados a imagens médicas (Kim *et al.*, 2019). Exemplos do uso das CNNs incluem a detecção de doenças por meio da análise de imagens de raio-X, conforme apresentado por Rajpurkar *et al.* (2017) e Qin *et al.* (2018).

Apesar dos benefícios, Habehh e Gohel (2021) apontam um dos riscos associados ao uso dessas tecnologias na área da saúde:

[...] é a disponibilidade de dados de formação e teste de alta qualidade com tamanhos de amostra suficientemente grandes para garantir elevada fiabilidade e reprodutibilidade das previsões. Dado que as abordagens baseadas no ML e no *deep learning* “aprendem” com os dados, nunca é demais sublinhar a importância dos dados de qualidade. (Habehh; Gohel, 2021, p. 296, tradução nossa)

3.3 Qualidade de Dados

A qualidade dos dados (DQ) “indica o grau em que os dados são completos, consistentes, livres de duplicação, precisos e oportunos para uma determinada finalidade.” (Gudivada; Apon; Ding, 2017, p.1, tradução nossa). A integridade dos dados é crucial em diversos campos, especialmente em *Big Data*, *Machine Learning* e *Deep Learning* devido à sua dependência de grandes volumes de dados para extrair valor e fazer previsões mais precisas. De acordo com Becker, King e McMullen (2015), o aumento do volume de dados é diretamente proporcional ao aumento dos problemas de qualidade dos dados.

Detectar e corrigir questões associadas aos dados é desafiador por causa da sua complexidade e à natureza específica de cada aplicação. A relevância dos problemas varia entre os domínios, assim como a distinção entre dados bons e ruins não é muito clara. Além disso, diversos tipos de problemas podem acontecer simultaneamente e os dados estão sempre mudando e alterando a natureza dos problemas ao longo do tempo (Dasu, 2013).

No contexto de *Big Data*, a qualidade dos dados é especialmente crítica devido ao volume, velocidade e variedade das informações. Manter a integridade dos dados em larga escala requer processos robustos de governança e tecnologias avançadas para identificação e correção de erros. Em *Deep Learning*, a dependência de grandes volumes de dados de alta qualidade é ainda mais acentuada. Modelos complexos de DL são altamente sensíveis a dados (Munappy *et al.*, 2022). Portanto, a qualidade dos dados deve ser cuidadosamente gerenciada

para garantir que os modelos possam aprender de forma eficaz e produzir resultados confiáveis.

Diferentes tipos de métodos devem ser aplicados para resolução de problemas como valores ausentes, duplicados, inconsistentes, atípicos ou até mesmo dados irrelevantes. Em *Machine Learning* podem ser aplicadas técnicas como pré-processamento de dados, limpeza e técnicas de aprimoramento (Soni *et al.*, 2023).

3.3.1 Dimensões de qualidade de dados

As dimensões da qualidade de dados (DQD) são um conjunto de critérios utilizados para medir a adequação e a qualidade dos dados para um determinado propósito. Cada dimensão de qualidade mede um atributo específico do dado (Mahanti, 2019). Segundo Ridzuan e Zainon (2024), as quatro principais dimensões são:

Acurácia: refere-se à exatidão com que os dados refletem a realidade, sendo confiável. Um dado é considerado preciso quando é exato e sem erros, ou seja, ele reflete a situação ou objeto que pretende descrever.

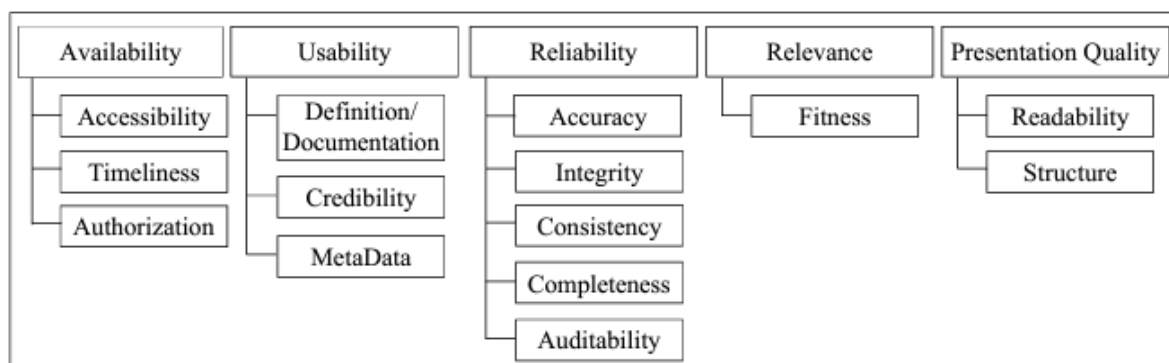
Consistência: assegura que os dados permaneçam uniformes em diferentes sistemas ou bases de dados. Dados consistentes não apresentam contradições entre si quando comparados ou integrados.

Completeness: Refere-se ao quanto os dados estão inteiramente presentes, sem lacunas. Dados completos são aqueles que contêm todas as informações necessárias para o uso pretendido, sem valores ausentes ou omissões importantes.

Temporalidade: Refere-se à atualidade e relevância temporal dos dados. Dados temporais são aqueles que estão dentro de um período de validade ou refletem o momento correto para o qual foram coletados.

Os autores ainda apontam que estas dimensões não são suficientes para medir a qualidade dos dados e que devem ser revisitadas, principalmente, para se adequar aos 5 V's do *Big Data*. Sendo assim, de forma a avaliar a qualidade dos dados de *Big Data*, Cai e Zhu (2015) propõem cinco dimensões (Figura 4): disponibilidade, usabilidade, confiabilidade, relevância e qualidade da apresentação, sendo esta última indicada pelos autores como adicional para melhora da satisfação do usuário.

Figura 4 - Um padrão universal de qualidade de *big data* de duas camadas para avaliação.



Fonte: Cai e Zhu (2015).

Cada dimensão possui de forma hierárquica elementos de qualidade de dados, que são aspectos mais específicos ou atributos que detalham a natureza daquela dimensão, e indicadores que são métricas específicas e mensuráveis, que quantificam os elementos de qualidade de dados.

4 METODOLOGIA

4.1 Revisão de Escopo

Revisão de escopo é um tipo de revisão de literatura que “é proposta para a realização de mapeamento da literatura num determinado campo de interesse, sobretudo quando revisões acerca do tema ainda não foram publicadas” (Cordeiro; Soares, 2019, p. 38) e que de acordo com Munn *et al.* (2018) são indicadas:

- Para identificar os tipos de evidências disponíveis em um determinado campo;
 - Para esclarecer os principais conceitos/definições na literatura;
 - Examinar como a pesquisa é conduzida em um determinado tópico ou campo;
 - Para identificar as principais características ou fatores relacionados a um conceito;
 - Como precursor de uma revisão sistemática;
 - Para identificar e analisar lacunas de conhecimento;
- (Munn *et al.*, 2018, p. 2, tradução nossa)

Para definição da metodologia aplicada ao estudo é importante entender a diferença entre a revisão de escopo e a revisão sistemática assim como a sua aplicação.

A revisão de escopo, diferente da revisão sistemática, não apresenta análises estatísticas por metanálise, medidas de sumarização, avaliação da qualidade dos estudos (avaliação do risco de viés) e outros métodos de análises adicionais. Adicionalmente, as revisões também se diferenciam quanto a sua aplicação, pois devem ser realizadas por motivos diferentes. As revisões sistemáticas são indicadas para confirmar se práticas atuais são baseadas em evidências, para comprovar a qualidade das evidências, para solucionar possíveis conflitos em evidências que possam estar induzindo a práticas discordantes, identificar lacunas e tendenciosidades nas evidências, para orientar a tomada de decisão e desenvolver políticas. (Coelho *et al.*, 2021, p.11)

Como o objetivo do trabalho em questão é entender os métodos utilizados na qualidade de dados de saúde no uso de Big Data associado a Machine Learning ou Deep Learning e os benefícios da aplicação destas tecnologias, optou-se pela realização de uma revisão de escopo seguindo o referencial metodológico proposto por Levac, Colquhoun e O'Brien (2010) em “*Scoping studies: advancing the methodology*”, que fornece recomendações para esclarecer e melhorar a metodologia de revisão de escopo apresentada por Arksey e O'Malley (2005). A estrutura tem seis etapas:

1. Identificação da questão de pesquisa.
2. Identificação de estudos relevantes.

3. Seleção dos estudos.
4. Mapeamento dos dados.
5. Agrupar, resumir e relatar os resultados.
6. Consulta (etapa opcional).

O *checklist* PRISMA-ScR (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews*), elaborado por Tricco *et al.* (2018), fornece diretrizes para realização de revisões de escopo, o qual foi utilizado para orientar e conduzir a realização de cada etapa da revisão. O *checklist* aplicado na pesquisa está disponível no Anexo A.

4.1.1 Identificação da questão de pesquisa

Esta pesquisa tem como objetivo determinar métodos e *frameworks* aplicados à qualidade de dados médicos e de saúde para *Big Data*, associados a *Machine Learning* ou *Deep Learning*. A partir do objetivo proposto, esta pesquisa visa responder às seguintes questões:

Quais são os benefícios da aplicação de *Big Data*, *Machine Learning* e *Deep Learning* na indústria da saúde? Como passo inicial no estudo da utilização de *Big Data*, ML e DL na saúde, é necessário compreender porque os autores optam pelo uso destas tecnologias.

Quais métodos são usados para qualidade de dados? Esta questão tem como objetivo identificar os métodos utilizados nos artigos e os possíveis benefícios do uso.

Que problemas estão associados à qualidade dos dados que este método atenua? A identificação dos problemas mitigados permite categorizar os desafios encontrados com esta tecnologia.

Quais dimensões de qualidade de dados de *Big Data* são tratadas para obter melhor qualidade dos dados? A identificação dessas dimensões apresentadas por Cai e Zhu (2015) possibilitam uma compreensão detalhada dos critérios usados para garantir que os dados sejam adequados ao seu propósito e cumpram os requisitos necessários para as tecnologias aplicadas.

4.1.2 Identificação dos estudos relevantes

Esta revisão foi realizada nas bases de dados Scopus Document Search, IEEE XploreDigital Library e ACM Digital Library, que são bases de dados multidisciplinares com uma vasta quantidade de trabalhos relacionados ao escopo desta revisão.

Foi realizada uma série de pesquisas por estudos primários relevantes para o tema com diferentes termos associados à qualidade de dados, *big data*, *machine learning*, *deep learning* e saúde, e a avaliação da relevância dos artigos obtidos. Após cinco rodadas de utilização de diferentes combinações de termos em português, inglês e espanhol, apenas foram obtidos artigos na utilização dos termos em inglês. Desta forma, a string final foi definida usando os termos (("*machine learning*" OR "*deep learning*") AND ("*big data*") AND ("*clean**" OR "*quality*" OR "*dirty*") AND ("*criteria*" OR "*process*" OR "*method*" OR "*model*") AND ("*medic**")).

4.1.3 Seleção dos estudos

Para determinar os critérios de elegibilidade dos estudos, foram discutidos os critérios de inclusão e exclusão para orientar a busca e revisão dos artigos obtidos. Os critérios de inclusão (INC) e exclusão (EXC) foram:

INC.01 - Domínio: O artigo deve abordar critérios, processos, métodos ou modelos aplicados à qualidade de dados de saúde para *Big Data* associado a *Machine Learning* ou *Deep Learning*.

INC.02 - Data: O artigo deve ter sido publicado nos últimos 5 anos.

INC.03 - Desenho do estudo: O artigo apresenta um estudo primário.

INC.04 - Idioma: O artigo deve estar escrito em Português, Inglês ou Espanhol.

INC.04 - Disponibilidade: O artigo completo deve estar disponível ou que permita o acesso através do Portal de Periódicos da CAPES¹ (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior).

EXC.01 - Duplicidade: O mesmo artigo é encontrado em outra base de dados. Neste cenário, as duplicatas são removidas.

EXC.02 - Não atende a todos os critérios de inclusão.

¹ <https://www.periodicos.capes.gov.br/>

EXC.03 - Retratado: O artigo foi retratado por descoberta de erros graves ou problemas éticos que comprometem a validade ou a integridade da pesquisa.

Com todos os artigos obtidos das buscas nas bases de dados e os critérios de elegibilidade definidos, foram realizadas as seguintes etapas:

1. Remoção de artigos duplicados, que é feita após avaliação dos artigos identificados como duplicados após a importação dos resultados da pesquisa para o Rayyan², software de gerenciamento de revisões sistemáticas.
2. Avaliar título, resumo, ano de publicação e palavras-chaves para retirada dos estudos que se enquadram nos critérios de exclusão.
3. Realizar revisão do texto completo, excluindo estudos que não atendam aos critérios e explicando os motivos da exclusão.

Nos casos de conflito nas etapas, os estudos foram discutidos com outros pesquisadores para encontrar a solução. Os resultados obtidos após seguir todas as etapas completaram o fluxograma PRISMA, que mostra a quantidade de artigos identificados, duplicados, triados e incluídos na revisão.

4.1.4 Mapeamento dos dados

Após a seleção dos estudos, os dados extraídos e tabulados dos artigos que atenderam aos critérios de inclusão foram inseridos em um formulário de dados que foi atualizado continuamente à medida que novos dados importantes foram identificados (Levac; Colquhoun; O'Brien, 2010). O formulário inclui:

- Título do artigo
- Autor(es)
- País
- Ano
- Editor/Periódico
- Objetivos do estudo
- Problemas e lacunas

² <https://www.rayyan.ai/>

- Métodos de qualidade de dados
- Benefícios dos métodos
- Problemas mitigados pelo método
- Dimensões de qualidade de dados
- Conclusão sobre o artigo

4.1.5 Agrupar, resumir e relatar os resultados

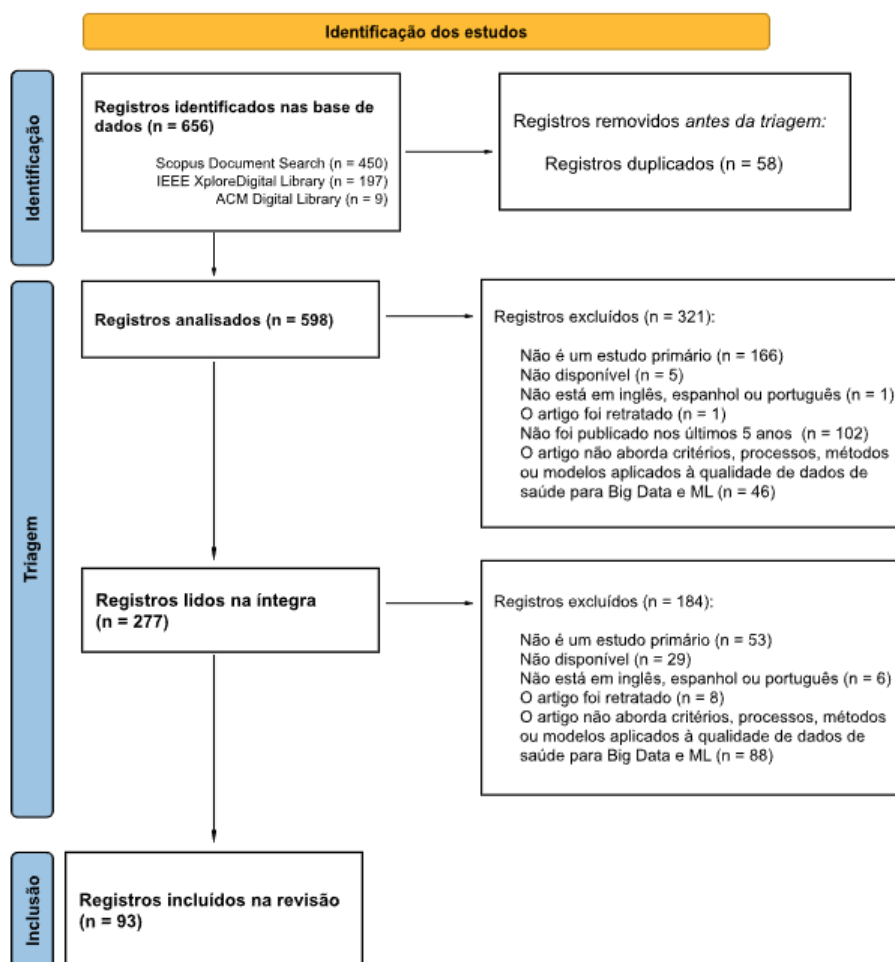
Conforme definido no framework de Levac, Colquhoun e O'Brien (2010), esta fase compreende as etapas de análise, relato dos resultados e aplicação de significado aos resultados. A análise das informações extraídas dos artigos foi realizada de forma qualitativa e quantitativa. A análise qualitativa permitiu o entendimento das técnicas aplicadas e das novas oportunidades de estudo. Na quantitativa, foram classificados e mensurados os métodos e técnicas mais utilizados e mapeados de acordo com as suas características, benefícios e os problemas mitigados.

Os resultados obtidos foram apresentados por meio de tabelas e gráficos, sendo estruturados e combinados com a análise realizada, servindo como base para a criação deste estudo. Além disso, foram discutidas as possíveis implicações dos resultados em pesquisas, práticas e políticas futuras. Os resultados serão debatidos com cientistas de dados, desenvolvedores e profissionais de saúde, buscando ampliar suas aplicações e impacto.

5 RESULTADOS

O processo de busca nas bases de dados retornou 656 artigos, que foram importados para o software Rayyan, o qual identificou possíveis duplicatas que precisam de revisão. Após essa etapa, constatou-se que 58 artigos estavam de fato duplicados. A triagem foi dividida em duas etapas, nas quais 598 artigos foram analisados de acordo com os critérios de inclusão e exclusão definidos na pesquisa. Na primeira etapa, foram avaliadas informações como título, resumo, ano de publicação e palavras-chave. Os 277 artigos que não foram excluídos com base nessas informações passaram para a segunda etapa de triagem, que consistiu na leitura completa do artigo. Após as etapas de triagem, 98 artigos foram incluídos para a revisão. Resultados quantitativos das etapas de seleção dos estudos estão apresentados no fluxograma PRISMA (Figura 5).

Figura 5 - Fluxograma PRISMA das etapas de seleção dos artigos.



Fonte: Adaptado de Page *et al.* (2021).

Os artigos incluídos na revisão são apresentados na Tabela 1, onde cada estudo é associado a um código de identificação para facilitar sua referência posterior, além de incluir o título, referência bibliográfica e periódico. Observa-se que, dos 93 registros analisados, 18,3% são de 2020, 28% de 2021, 30,1% de 2022, 22,6% de 2023 e 1,1% de 2024.

Tabela 1 - Estudos incluídos na revisão de escopo.

Id	Título	Referência Bibliográfica	Periódico
A1	<i>Ambient context-based modeling for health risk assessment using deep neural network</i>	CHUNG, Kyungyong; YOO, Hyun; CHOE, Do-Eun (2020)	<i>Journal of Ambient Intelligence and Humanized Computing</i>
A2	<i>Development and Application of an Intensive Care Medical Data Set for Deep Learning</i>	ZHAO, Shangping <i>et al.</i> (2020)	<i>Proceedings - 2020 IEEE International Conference on Big Data, Big Data 2020</i>
A3	<i>Modeling of Clinical Mammography Recognition</i>	CHU, Kuo-Chung <i>et al.</i> (2020)	<i>Proceedings - 2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science, IRI 2020</i>
A4	<i>Heterogeneous Similarity Graph Neural Network on Electronic Health Records</i>	LIU, Zheng <i>et al.</i> (2020)	<i>Proceedings - 2020 IEEE International Conference on Big Data, Big Data 2020</i>
A5	<i>Performance evaluation of a proposed machine learning model for chronic disease datasets using an integrated attribute evaluator and an improved decision tree classifier</i>	MISHRA, Sushruta <i>et al.</i> (2020)	<i>Applied Sciences (Switzerland)</i>
A6	<i>Study on the related factors of post-herpetic neuralgia in hospitalized patients with herpes zoster in Sichuan Hospital of Traditional Chinese Medicine based on big data analysis</i>	LI, Tianhao <i>et al.</i> (2020)	<i>Dermatologic Therapy</i>
A7	<i>A systematic quality assurance framework for the upgrade of radiation oncology information systems</i>	ZHANG, Baoshe <i>et al.</i> (2020)	<i>Physica Medica</i>

A8	<i>Integrated natural language processing and machine learning models for standardizing radiotherapy structure names</i>	SYED, Khajamoinuddin <i>et al.</i> (2020)	<i>Healthcare (Switzerland)</i>
A9	<i>A Qualitative Assessment of Machine Learning Support for Detecting Data Completeness and Accuracy Issues to Improve Data Analytics in Big Data for the Healthcare Industry</i>	JUDDOO, Suraj; GEORGE, Carlisle (2020)	<i>2020 3rd International Conference on Emerging Trends in Electrical, Electronic and Communications Engineering (ELECOM)</i>
A10	<i>Differential Privacy Practice on Diagnosis of COVID-19 Radiology Imaging Using EfficientNet</i>	MÜFTÜOĞLU, Zümrit; KIZRAK, M. Ayyüce; YILDLNM, Tülay (2020)	<i>2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)</i>
A11	<i>Predicting Asthma Attacks: Effects of Indoor PM Concentrations on Peak Expiratory Flow Rates of Asthmatic Children</i>	KIM, Dohyeong <i>et al.</i> (2020)	<i>IEEE Access</i>
A12	<i>Spatially-Constrained Fisher Representation for Brain Disease Identification With Incomplete Multi-Modal Neuroimages</i>	PAN, Yongsheng <i>et al.</i> (2020)	<i>IEEE Transactions on Medical Imaging</i>
A13	<i>A Preliminary Study on Projection Denoising for Low-dose CT Imaging Using Modified Dual-Domain U-net</i>	FENG, Zhiwei <i>et al.</i> (2020)	<i>2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD)</i>
A14	<i>Comparative Analysis of Pre-process Pipelines For Automatic Retinal Vessel Segmentation</i>	LEI, Gaoyi <i>et al.</i> (2020)	<i>2020 39th Chinese Control Conference (CCC)</i>
A15	<i>Mining Domain Knowledge: Improved Framework Towards Automatically Standardizing Anatomical Structure Nomenclature in Radiotherapy</i>	YANG, Qiming <i>et al.</i> (2020)	<i>IEEE Access</i>
A16	<i>Features Importance to Improve Interpretability of Chronic Kidney Disease Early Diagnosis</i>	MORENO-SANCHEZ, Pedro A. (2020)	<i>2020 IEEE International Conference on Big Data (Big Data)</i>
A17	<i>Automated Semantic Segmentation of Red Blood Cells for Sickle Cell Disease</i>	ZHANG, Mo <i>et al.</i> (2020)	<i>IEEE Journal of Biomedical and Health Informatics</i>

A18	<i>Research on Design of Fog Computing Optimization Model for Medical Big Data</i>	QIN, Baoling. (2021)	<i>Advances in Transdisciplinary Engineering</i>
A19	<i>Analyzing groups of inpatients' healthcare needs to improve service quality and sustainability</i>	HSU, Ming-Hsia <i>et al.</i> (2021)	<i>Sustainability (Switzerland)</i>
A20	<i>Sleep stage classification using extreme learning machine and particle swarm optimization for healthcare big data</i>	SURANTHA, Nico; LESMANA, Tri Fennia; ISA, Sani Muhamad (2021)	<i>Journal of Big Data</i>
A21	<i>Development of a predictive model for integrated medical and long-term care resource consumption based on health behaviour: application of healthcare big data of patients with circulatory diseases</i>	TAKURA, Tomoyuki; HIRANO GOTO, Keiko; HONDA, Asao (2021)	<i>BMC Medicine</i>
A22	<i>Computer Audition for Fighting the SARS-CoV-2 Corona Crisis-Introducing the Multitask Speech Corpus for COVID-19</i>	QIAN, Kun <i>et al.</i> (2021)	<i>IEEE Internet of Things Journal</i>
A23	<i>MRI Image Reconstruction Based on Artificial Intelligence</i>	LIANG, NaNa (2021)	<i>Journal of Physics: Conference Series</i>
A24	<i>A deep neural network-based method for prediction of dementia using big data</i>	KIM, Jungyoon; LIM, Jihye (2021)	<i>International Journal of Environmental Research and Public Health</i>
A25	<i>Application of U-Net with Global Convolution Network Module in Computer-Aided Tongue Diagnosis</i>	LI, Meng-Yi <i>et al.</i> (2021)	<i>Journal of Healthcare Engineering</i>
A26	<i>Trading-off machine learning algorithms towards data-driven administrative-socio-economic population health management</i>	PANICACCI, Silvia <i>et al.</i> (2021)	<i>Computers</i>
A27	<i>Social mining-based clustering process for big-data integration</i>	JUNG, Hoill; CHUNG, Kyungyong (2021)	<i>Journal of Ambient Intelligence and Humanized Computing</i>
A28	<i>The Temple University Digital Pathology Corpus: The Breast Tissue Subset</i>	WEVODAU, Z. <i>et al.</i> (2021)	<i>2021 IEEE Signal Processing in Medicine and Biology Symposium, SPMB 2021 - Proceedings</i>

A29	<i>A Data-driven Method for BioBrick Quality Assessment</i>	AN, Ning <i>et al.</i> (2021)	<i>2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)</i>
A30	<i>Assessing YOLACT++ for real time and robust instance segmentation of medical instruments in endoscopic procedures</i>	CERÓN, Juan Carlos Angeles <i>et al.</i> (2021)	<i>2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)</i>
A31	<i>A Deep Learning-Based Automatic First-Arrival Picking Method for Ultrasound Sound-Speed Tomography</i>	QU, Xiaolei <i>et al.</i> (2021)	<i>IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control</i>
A32	<i>3D Deep Convolutional Neural Network for Detection of Anomalous Rigid and Affine Registrations in Big Data Brain MRI</i>	TUMMALA, Jagadeesh <i>et al.</i> (2021)	<i>2021 IEEE Bombay Section Signature Conference (IBSSC)</i>
A33	<i>Early Stage Detection of Scoliosis Using Machine Learning Algorithms</i>	SHRESTHA, Pratik <i>et al.</i> (2021)	<i>2021 International Conference on Forensics, Analytics, Big Data, Security (FABS)</i>
A34	<i>An Ontology-Driven Structured Processing Method for History of Present Illness of Discharge Summary</i>	RUI, Xinkai <i>et al.</i> (2021)	<i>2021 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCCom/CyberSciTech)</i>
A35	<i>Applying Cross-Modality Data Processing for Infarction Learning in Medical Internet of Things</i>	XU, Chenchu <i>et al.</i> (2021)	<i>IEEE Internet of Things Journal</i>
A36	<i>High-Risk Prediction of Cardiovascular Diseases via Attention-Based Deep Neural Networks</i>	AN, Ying <i>et al.</i> (2021)	<i>IEEE/ACM Transactions on Computational Biology and Bioinformatics</i>

A37	<i>Detection algorithm of pulmonary nodules based on deep learning</i>	ZHANG, Weiguo; CUI, Linfang (2021)	<i>2021 2nd International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)</i>
A38	<i>Improving Brain Dysfunction Prediction by GAN: A Functional-Connectivity Generator Approach</i>	YAN, Da <i>et al.</i> (2021)	<i>2021 IEEE International Conference on Big Data (Big Data)</i>
A39	<i>Diabetic Retinopathy Diagnosis Using Multichannel Generative Adversarial Network With Semisupervision</i>	WANG, Shuqiang <i>et al.</i> (2021)	<i>IEEE Transactions on Automation Science and Engineering</i>
A40	<i>Examination Of Diabetes Mellitus For Ahead Of Schedule Expectation Utilizing Ideal Highlights Determination</i>	KUMAR, M. Jogendra <i>et al.</i> (2021)	<i>2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)</i>
A41	<i>A Dual Weighted Shared Capsule Network for Diabetic Retinopathy Fundus Classification</i>	NNEJI, Grace Ugochi <i>et al.</i> (2021)	<i>2021 International Conference on High Performance Big Data and Intelligent Systems (HPBD&IS)</i>
A42	<i>A New Tele-Healthcare System of Heart Sound Signal Classification based on Gated Recurrent Unit</i>	OKTARINA, Eka Sari; SETIAWAN, Noor Akhmad; ARDIYANTO, Igi (2021)	<i>2021 International Conference on Artificial Intelligence and Big Data Analytics</i>
A43	<i>Noise Reduction for SD-OCT Using a Structure-Preserving Domain Transfer Approach</i>	WU, Menglin <i>et al.</i> (2021)	<i>IEEE Journal of Biomedical and Health Informatics</i>
A44	<i>Federated learning enables big data for rare cancer boundary detection</i>	PATI, Sarthak <i>et al.</i> (2022)	<i>Nature Communications</i>
A45	<i>Smart health analysis system using regression analysis with iterative hashing for IoT communication networks</i>	REHMAN, Amjad <i>et al.</i> (2022)	<i>Computers and Electrical Engineering</i>

A46	<i>Disease-specific data processing: An intelligent digital platform for diabetes based on model prediction and data analysis utilizing big data technology</i>	KONG, Xiangyong <i>et al.</i> (2022)	<i>Frontiers in Public Health</i>
A47	<i>A robust framework for handling health care information based on machine learning and big data engineering techniques</i>	PRAVEEN, S. Phani <i>et al.</i> (2022)	<i>International Journal of Healthcare Management</i>
A48	<i>Fall Detection Scheme based on Deep Learning Model for High-Quality Life</i>	JAYAKARTHIK, R. <i>et al.</i> (2022)	<i>3rd International Conference on Electronics and Sustainable Communication Systems, ICESC 2022 - Proceedings</i>
A49	<i>Research on Key Technologies of Multi-modal Heterogeneous Medical Big Data Based on Deep Learning</i>	ZENG, Yan; LIANG, Wei (2022)	<i>ISCTT 2022 - 7th International Conference on Information Science, Computer Technology and Transportation</i>
A50	<i>Automatic Control System for Quality Assurance of Clinical Pathology Examination: A Machine Learning Approach</i>	OH, J. W.; YOO, Og-Woo; KANG, Jin-Kyu (2022)	<i>Journal of System and Management Sciences</i>
A51	<i>Extended and optimized deep convolutional neural network-based lung tumor identification in big data</i>	ANANTH, Antony Dennis; PALANISAMY, Chenniappan (2022)	<i>International Journal of Imaging Systems and Technology</i>
A52	<i>Study on the Relationship between Lung Cancer Stromal Cells and Air Cavity Diffusion Based on an Image Acquisition System</i>	BAI, Shan <i>et al.</i> (2022)	<i>Contrast Media and Molecular Imaging</i>
A53	<i>Class Imbalance Handling with Deep Learning Enabled IoT Healthcare Diagnosis Model</i>	RAGUPATHI, T.; GOVINDARAJAN, M.; PRIYARADHIKADEVI, T. (2022)	<i>Intelligent Automation and Soft Computing</i>
A54	<i>A Healthcare Quality Assessment Model Based on Outlier Detection Algorithm</i>	ALHARBE, Nawaf; RAKROUKI, Mohamed Ali; ALJOHANI, Abeer. (2022)	<i>Processes</i>

A55	<i>A Study of the Automatic Authentication Module to Improve the Consistency of Outsourced Laboratory Test Results</i>	CHOI, Jin.-Il.; OH, Jai.-Woo (2022)	<i>Journal of System and Management Sciences</i>
A56	<i>A Modelling Framework for Evidence-Based Public Health Policy Making</i>	PRASINOS, Marios <i>et al.</i> (2022)	<i>IEEE Journal of Biomedical and Health Informatics</i>
A57	<i>Deep enriched salp swarm optimization based bidirectional -long short term memory model for healthcare monitoring system in big data</i>	DHAND, Geetika <i>et al.</i> (2022)	<i>Informatics in Medicine Unlocked</i>
A58	<i>Medical Big Data Analysis using Binary Moth-Flame with Whale Optimization Approach</i>	RAO, Saka Uma Maheswara; RAO, K. Venkata; PVGD, Prasad Reddy. (2022a)	<i>International Journal of Advanced Computer Science and Applications</i>
A59	<i>Enhanced Efficient and Security in Big Data Using TDES And Machine Learning Technique</i>	SHRIHARI, M. R. <i>et al.</i> (2022)	<i>IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics, ICDCECE 2022</i>
A60	<i>Deep learning Based Patient-Friendly Clinical Expert Recommendation Framework</i>	KUMAR, Akhilesh <i>et al.</i> (2022)	<i>Proceedings of 2nd International Conference on Innovative Practices in Technology and Management, ICIPTM 2022</i>
A61	<i>Medical Big Data Analysis using LSTM based Co-Learning Model with Whale Optimization Approach</i>	RAO, Saka Uma Maheswara; RAO, K. Venkata; PVGD, Prasad Reddy. (2022b).	<i>International Journal of Intelligent Engineering and Systems</i>
A62	<i>Quality prediction through machine learning for the inspection and manufacturing process of blood glucose test strips</i>	TSOU, Ching-Shih <i>et al.</i> (2022)	<i>Cogent Engineering</i>
A63	<i>Automated Detection of Covid-19 from X-ray Using SVM</i>	ZEBARI, Dilovan Asaad <i>et al.</i> (2022)	<i>ICOASE 2022 - 4th International Conference on Advanced Science and Engineering</i>

A64	<i>Prediction of complications of type 2 Diabetes: A Machine learning approach</i>	NICOLUCCI, Antonio <i>et al.</i> (2022)	<i>Diabetes Research and Clinical Practice</i>
A65	<i>Concept and Proof of the Lifelog Bigdata Platform for Digital Healthcare and Precision Medicine on the Cloud</i>	LEE, Kyu Hee <i>et al.</i> (2022)	<i>Yonsei Medical Journal</i>
A66	<i>Prediction of Tumor in Mammogram Images Using Data Mining Models</i>	SARRAY, Noor; INAN, Timur. (2022)	<i>HORA 2022 - 4th International Congress on Human-Computer Interaction, Optimization and Robotic Applications, Proceedings</i>
A67	<i>Cough Classification with Deep Derived Features using Audio Spectrogram Transformer</i>	VALDÉS, Julio <i>et al.</i> (2022)	<i>2022 IEEE International Conference on Big Data (Big Data)</i>
A68	<i>Health Insurance Cost Prediction Using Regression Models</i>	PANDA, Sudhir <i>et al.</i> (2022)	<i>2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)</i>
A69	<i>Learning COVID-19 Pneumonia Lesion Segmentation From Imperfect Annotations via Divergence-Aware Selective Training</i>	YANG, Shuojue <i>et al.</i> (2022)	<i>IEEE Journal of Biomedical and Health Informatics</i>
A70	<i>A Diabetes Prediction System Based on Federated Learning</i>	LIU, Junzhe <i>et al.</i> (2022)	<i>2022 International Conference on Big Data, Information and Computer Network (BDICN)</i>
A71	<i>ChroSegNet: An Attention-Based Model for Chromosome Instance Segmentation</i>	CHEN, Xiaoyu; CAI, Qiang; CHEN, Xiaoming. (2022)	<i>HBDSS 2022; 2nd International Conference on Health Big Data and Smart Sports</i>
A72	<i>Data-Centric AI for Healthcare Fraud Detection</i>	JOHNSON, Justin M.; KHOSHGOFTAAR, Taghi M. (2023)	<i>SN Computer Science</i>
A73	<i>Diabetes Monitoring System in Smart Health Cities Based on Big Data Intelligence</i>	ALZU'BI, Shadi <i>et al.</i> (2023)	<i>Future Internet</i>

A74	<i>Improvement of Dialysis Dosing Using Big Data Analytics</i>	MUMTAZ, Syeda Leena <i>et al.</i> (2023)	<i>Healthcare Informatics Research</i>
A75	<i>A deep learning approach to private data sharing of medical images using conditional generative adversarial networks (GANs)</i>	SUN, Hanxi <i>et al.</i> (2023)	<i>PLoS ONE</i>
A76	<i>An ECG denoising method based on adversarial denoising convolutional neural network</i>	HOU, Yanrong <i>et al.</i> (2023)	<i>Biomedical Signal Processing and Control</i>
A77	<i>Trustworthy Privacy-Preserving Hierarchical Ensemble and Federated Learning in Healthcare 4.0 With Blockchain</i>	STEPHANIE, Veronika <i>et al.</i> (2022)	<i>IEEE Transactions on Industrial Informatics</i>
A78	<i>Optimized U-Net convolutional neural network based breast cancer prediction for accuracy increment in big data</i>	KIROLA, Madhu; MEMORIA, Minakshi; DUMKA, Ankur (2023)	<i>Concurrency and Computation: Practice and Experience</i>
A79	<i>Using machine learning to improve anaphylaxis case identification in medical claims data</i>	KURAL, Kamil Can <i>et al.</i> (2023)	<i>JAMIA Open</i>
A80	<i>A novel variant of deep convolutional neural network for classification of ovarian tumors using CT images</i>	KODIPALLI, Ashwini <i>et al.</i> (2023)	<i>Computers and Electrical Engineering</i>
A81	<i>Artificial intelligence-based mining of electronic health record data to accelerate the digital transformation of the national cardiovascular ecosystem: Design protocol of the CardioMining study</i>	SAMARAS, Athanasios <i>et al.</i> (2023)	<i>BMJ Open</i>
A82	<i>Hierarchical Hybrid Networks for Automatic Pulmonary Blood Vessel Segmentation in Computed Tomography Images</i>	ZHOU, Qingguo <i>et al.</i> (2023)	<i>IEEE/ACM Transactions on Computational Biology and Bioinformatics</i>
A83	<i>Secure Image Retrieval of Poor Quality Images by Combining LE-GAN, Arnold Mapping and Logistic Mapping</i>	MOHAN, Maya <i>et al.</i> (2023)	<i>2023 4th International Conference for Emerging Technology, INCET 2023</i>

A84	<i>Timely Prediction of Diabetes by Means of Machine Learning Practices</i>	TRIPATHI, Rajan Prasad <i>et al.</i> (2023)	<i>Augmented Human Research</i>
A85	<i>Metadata-independent classification of MRI sequences using convolutional neural networks: Successful application to prostate MRI</i>	BAUMGÄRTNER, Georg L. <i>et al.</i> (2023)	<i>European Journal of Radiology</i>
A86	<i>Big Data-based Smart Health Monitoring System: Using Deep Ensemble Learning</i>	ABIDI, Mustufa Haider <i>et al.</i> (2023)	<i>IEEE Access</i>
A87	<i>The Precision in Psychiatry (PIP) study: Testing an internet-based methodology for accelerating research in treatment prediction and personalisation</i>	LEE, Chi Tak <i>et al.</i> (2023)	<i>BMC Psychiatry</i>
A88	<i>Measuring Sleep Stages and Quality Based on K-Means and Random Forest Algorithms</i>	LI, Jiahui <i>et al.</i> (2023)	<i>2023 4th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT)</i>
A89	<i>U-Shaped Network with Multi-Scale Feature Extraction for Lesion Region Segmentation in Gastric Cancer Images</i>	ZHANG, Yuanyu; LI, Yunqi; HE, Dongzhi. (2023)	<i>2023 IEEE 3rd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA)</i>
A90	<i>Remote photoplethysmography signals enhancement based on generative adversarial networks</i>	KUANG, Hailan <i>et al.</i> (2023)	<i>2023 IEEE 3rd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA)</i>
A91	<i>Multilevel Stochastic Optimization for Imputation in Massive Medical Data Records</i>	LI, Wenrui <i>et al.</i> (2023)	<i>IEEE Transactions on Big Data</i>
A92	<i>Breast Cancer Classification Based on DNA Microarray Analysis</i>	EL RAHMAN, Sahar A.; ALLUHAIDAN, Ala Saleh; MARZOUK, Radwa (2023)	<i>IEEE Access</i>

A93	<i>Efficient and secure medical big data management system using optimal map-reduce framework and deep learning</i>	RAJESHKUMAR, K.; DHANASEKARAN, S.; VASUDEVAN, V. (2024).	<i>Multimedia Tools and Applications</i>
------------	---	--	--

Fonte: Elaborada pelo autor (2024).

A partir dos artigos inseridos na revisão podemos responder as questões da pesquisa:

Quais são os benefícios da aplicação de Big Data, Machine Learning e Deep Learning na indústria da saúde?

Os autores abordam que a necessidade do uso de *Machine Learning* e *Deep Learning* associados ao *Big Data* é essencial devido ao grau de complexidade na análise de grandes volumes de dados (A7, A16, A20, A28, A29, A46, A51, A57, A68, A81, A86, A91), à melhoria na precisão dos resultados (A10, A14, A16, A32, A33, A37, A38, A40, A44, A48, A49, A50, A58, A61, A66, A76, A78, A79, A84, A86, A89, A91), à redução no tempo de resposta (A2, A23, A30, A33, A38, A48, A66, A87) e à de tomada de decisão orientada por dados (A4, A11, A19, A24, A25, A47, A55, A56, A60, A74, A88). Em contrapartida, um dos principais desafios apontados que podem afetar a qualidade do uso destas tecnologias na indústria da saúde são os associados à qualidade dos dados utilizados.

Quais métodos são usados para qualidade de dados? Que problemas estão associados à qualidade dos dados que este método atenua?

De forma a mitigar os problemas derivados da má qualidade dos dados em *Big Data*, ML e DL, diferentes métodos são aplicados antes do processamento ou análise dos dados. Na Tabela 2, são apresentados os diferentes métodos utilizados, seus benefícios, problemas mitigados e os estudos no qual foram aplicados.

Tabela 2 - Análise dos métodos de qualidade de dados, benefícios e problemas mitigados.

Categoria	Método	Benefícios	Problema Mitigados	Artigos
Métodos de Limpeza e Transformação	Limpeza de Dados: Remoção de dados inconsistentes, inválidos, duplicados, incompletos ou ruidosos.	Redução de erros, aumento da precisão, melhor interpretabilidade e consistência dos dados.	Dados inconsistentes, dados incorretos, dados duplicados ou ruidos.	A1, A2, A3, A4, A6, A7, A9, A11, A13, A18, A19, A20, A22, A26, A27, A28, A34, A36, A37, A42, A46, A48, A50, A54, A55, A56, A57, A58, A59, A61, A63, A68, A73, A76, A84, A85, A88, A89, A90, A92, A93
	Normalização: Ajuste dos dados para uma escala comum.	Aumento da precisão, melhor desempenho do modelo e comparabilidade entre variáveis.	Viés devido à inconsistência de dados ou escalas diferentes entre variáveis.	A1, A3, A11, A14, A15, A16, A18, A20, A23, A29, A32, A33, A37, A40, A59, A63, A74, A84, A88, A89
	Imputação de Valores Ausentes: Preenchimento de dados faltantes usando métodos como média, mediana, moda ou algoritmos mais complexos.	Aumento da completude dos dados, redução do impacto de dados faltantes na análise e maximização do uso dos dados disponíveis.	Dados faltantes ou dados incompletos.	A1, A2, A12, A16, A19, A26, A40, A54, A56, A57, A64, A68, A72, A74, A84, A87, A91, A93
	Remoção de <i>Outliers</i> : Identificação e remoção de valores extremos que podem distorcer a análise.	Aumento da precisão, melhor interpretabilidade e redução de erros.	Dados inconsistentes, outliers na segmentação.	A9, A19, A51, A54, A78, A84

	Transformação de Dados: Conversão de dados de um formato para outro.	Compatibilidade com modelos, melhor interpretabilidade, aumento da precisão, melhor eficiência de processamento.	Dados inconsistentes, dados categóricos, dificuldade na integração de dados.	A1, A19, A26, A37, A40, A42, A45, A47, A54, A55, A57, A63, A68, A74, A90, A92
	Balanceamento de Dados: Ajustar a distribuição das classes para evitar viés.	Melhora na precisão da predição, mitigação do impacto de dados desequilibrados, melhor desempenho geral do modelo.	Dados desequilibrados, viés.	A19, A26, A32, A44, A53, A62, A64
Métodos de Seleção e Extração de Características	Extração de Características: Seleção de atributos relevantes dos dados.	Redução da dimensionalidade, aumento da precisão, melhor desempenho do modelo, redução do tempo de processamento, melhor interpretabilidade.	Dados redundantes e ruidosos, atributos irrelevantes, redundância de informações, alto número de variáveis.	A1, A8, A11, A12, A14, A15, A16, A20, A21, A24, A29, A36, A39, A48, A49, A51, A52, A60, A63, A75, A78, A80, A83, A89, A92
	Seleção de Características/ Atributos: Escolha dos atributos mais relevantes para o modelo.	Redução do conjunto de atributos, aumento da precisão, redução do tempo de processamento, melhor interpretabilidade, melhor desempenho do modelo.	Dados redundantes e ruidosos, atributos irrelevantes.	A5, A8, A16, A19, A21, A26, A36, A40, A49, A57, A62, A72, A73, A78, A79, A84, A86, A92
Métodos Relacionados à Privacidade e Segurança	Criptografia: Proteção de dados sensíveis contra acesso não autorizado.	Segurança aprimorada, confidencialidade dos dados, integridade dos dados.	Riscos de segurança, acesso não autorizado, vulnerabilidades de segurança, falhas de segurança.	A45, A59, A70, A77, A83, A93

	Desidentificação / Anonimização: Remoção de informações que possam identificar indivíduos.	Proteção da privacidade do paciente, conformidade com regulamentações de privacidade.	Riscos de privacidade, dificuldades de anonimização.	A18, A65, A81, A82, A87
	Controle de Acesso: Restrição de acesso a dados com base em privilégios.	Segurança aprimorada, proteção contra acesso não autorizado.	Riscos de segurança, acesso não autorizado.	A45, A47, A59, A65, A70
	Privacidade Diferencial (DP): Adiciona ruído aos dados para proteger a privacidade individual.	Proteção da privacidade dos dados, permite treinar modelos com dados sensíveis.	Risco de vazamento de dados sensíveis.	A10
	Geração de Dados Sintéticos: Cria dados artificiais que preservam as características estatísticas dos dados originais, mas não contêm informações reais dos pacientes.	Permite o compartilhamento de dados sem comprometer a privacidade, podendo mitigar problemas de dados faltantes.	Risco de privacidade, limitação de dados.	A12, A38, A75, A77
Processamento Específico de Imagem	Recorte/ Mascaramento: Remoção de partes irrelevantes da imagem.	Melhora na qualidade da imagem, redução de erros na segmentação, foco nas áreas de interesse, redução da carga computacional.	Ruído e dados irrelevantes, influência de ruído e artefatos, sobrecarga computacional, distração do modelo por informações irrelevantes.	A13, A15, A17, A25, A31, A32, A37, A63, A75, A78, A80, A89

	Aumento de Dados: Criação de novas amostras a partir dos dados existentes.	Mitigação da escassez de dados, redução de overfitting, melhor generalização do modelo.	Escassez de dados, overfitting.	A13, A25, A30, A37, A38, A39, A48, A51, A63, A67, A69, A71, A76, A78, A89
	Fusão de Dados/Imagens: Combinação de dados de múltiplas imagens.	Aumento da usabilidade dos dados, visão mais abrangente dos dados, aumento da sensibilidade e especificidade do diagnóstico.	Dados inconsistentes, dados faltantes, baixa precisão diagnóstica.	A1, A12, A49, A50, A52
	Equalização de Histograma: Ajusta o contraste da imagem.	Aumento do contraste, melhor visualização das características.	Ruído e artefatos, baixa qualidade da imagem.	A3, A14, A41, A63, A71, A78

Fonte: Elaborada pelo autor (2024).

Quais dimensões de qualidade de dados de *Big Data* são tratadas para obter melhor qualidade dos dados?

A Tabela 2 demonstra os métodos utilizados para o tratamento dos dados, com o objetivo de torná-los mais adequados para o seu uso. Já a Tabela 3, abaixo, apresenta uma análise sobre a aplicação direta ou indireta das DQD propostas por Cai e Zhu (2015) nos artigos incluídos na revisão. Embora os quadros abordam a qualidade dos dados sob diferentes perspectivas, ambos convergem para o objetivo comum de garantir que os dados utilizados sejam adequados para o seu propósito, confiáveis e capazes de gerar resultados significativos.

Tabela 3 - Análise das Dimensões de Qualidade de Dados.

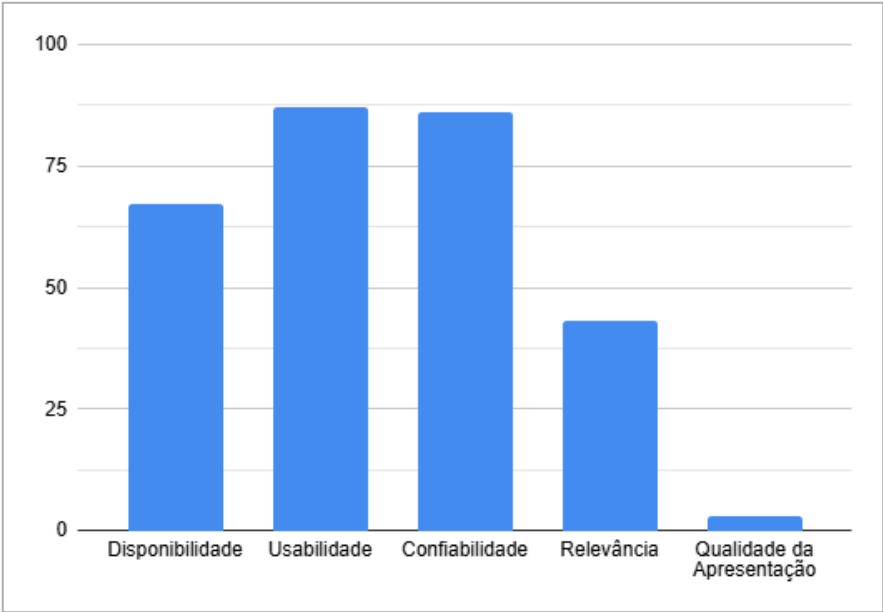
Dimensão	Artigos
Disponibilidade	A2, A4, A9, A10, A12, A14, A15, A16, A17, A18, A19, A21, A22, A26, A27, A29, A30, A32, A34, A35, A36, A37, A38, A39, A40, A41, A42, A43, A44, A45, A46, A47, A48, A49, A50, A51, A52, A54, A55, A56, A57, A58, A59, A60, A61, A64, A65, A67, A70, A71, A72, A74, A75, A76, A77, A78, A79, A81, A82, A84, A85, A86, A87, A89, A91, A92, A93

Usabilidade	A1, A2, A3, A4, A5, A6, A8, A9, A10, A11, A12, A13, A14, A15,A16, A17, A18, A19, A20, A21, A22, A23, A24, A25, A26, A27, A28, A29, A30, A31, A32, A33, A34, A35, A36, A37, A38, A39, A40, A41, A42, A43, A45, A46, A47, A49, A50, A51, A52, A54, A55, A56, A57, A58, A59, A60, A61, A64, A65, A66, A67, A68, A69, A70, A71, A72, A73, A74, A75, A76, A77, A78, A79, A80, A81, A82, A83, A84, A85, A86, A87, A88, A89, A90, A91, A92, A93
Confiabilidade	A1, A2, A4, A5, A6, A7, A8, A9, A10, A11, A12, A13, A14, A15, A16, A17, A18, A19, A20, A21, A22, A23, A24, A25, A26, A27, A28, A29, A30, A31, A32, A33, A34, A35, A36, A37, A38, A39, A40, A41, A42, A43, A44, A45, A46, A47, A48, A49, A50, A51, A52, A54, A55, A56, A57, A58, A59, A60, A63, A64, A65, A66, A67, A68, A69, A70, A71, A72, A73, A74, A75, A77, A78, A79, A80, A81, A82, A83, A85, A86, A87, A88, A89, A90, A91, A92
Relevância	A1, A2, A6, A8, A10, A11, A12,A16, A17, A18, A21, A23, A24, A26, A28, A30, A34, A40, A44, A46, A47, A50, A53, A57, A59, A60, A61, A62, A65, A67, A72, A73, A74, A75, A79, A82, A84, A86, A87, A89, A90, A92, A93
Qualidade da Apresentação	A1, A50, A87

Fonte: Elaborada pelo autor (2024).

A partir da análise da Tabela 3 foi observada a frequência com que cada dimensão de qualidade de dados é abordada nos artigos, conforme Gráfico 1.

Gráfico 1 - Quantitativo de estudos por Dimensão de Qualidade de Dados.



Fonte: Elaborada pelo autor (2024).

A usabilidade é a dimensão predominante, sendo tratada em 87 dos 93 artigos, reforçando a importância de garantir que os dados sejam úteis e adequados às necessidades dos usuários. Logo após a confiabilidade, é discutida em 86 artigos, demonstrando que é necessária a confiança nos dados. A disponibilidade, que avalia a facilidade do acesso aos dados, aparece em 67 artigos, enquanto a relevância, que trata da adequação e conformidade dos dados aos objetivos da pesquisa, é observada em 43 artigos. A dimensão qualidade da apresentação, que se refere à clareza na apresentação dos dados, é a menos frequente, aparecendo em apenas 3 artigos. Essa diferença na frequência indica que, embora todas as dimensões sejam importantes para a qualidade dos dados, a ênfase das pesquisas se concentra em garantir a usabilidade e confiabilidade, enquanto a qualidade da apresentação é considerada um aspecto secundário.

6 DISCUSSÃO

A partir da revisão de escopo foram encontrados 93 artigos cuja data de publicação datam entre 2020 e 2024. Os estudos evidenciaram o sucesso de projetos de *Big Data*, *Machine Learning* e *Deep Learning* na indústria da saúde, fortalecendo ainda mais o avanço rápido do setor (Chakraborty *et al.*, 2024). Os resultados obtidos nesta revisão permitiram identificar os principais métodos de pré-processamento utilizados para obter melhor qualidade dos dados.

A análise revelou que técnicas como limpeza, transformação e integração de dados são amplamente aplicadas para mitigar problemas comuns, como inconsistências, duplicidade e valores ausentes, que comprometem a eficácia dos sistemas baseados em dados na saúde. A principal contribuição deste trabalho reside no mapeamento detalhado de métodos de qualidade aplicados a dados de saúde em contextos de *Big Data*, ML e DL. Este estudo não apenas preenche lacunas na literatura ao sistematizar essas práticas, mas também oferece uma base teórica para que cientistas de dados, desenvolvedores e gestores em saúde adotem estratégias eficazes na manipulação e análise de grandes volumes de dados.

O estudo também destacou a importância das dimensões de qualidade de dados. As DQD permitem metrificar e avaliar a disponibilidade, usabilidade, confiabilidade, relevância e qualidade da apresentação dos dados, o que influencia no desempenho dos sistemas. A usabilidade, confiabilidade e disponibilidade foram as dimensões mais utilizadas para medir a qualidade dos dados. A relevância, que é uma das quatro dimensões indispensáveis na qualidade dos dados, é abordada em apenas 43 artigos, desta forma se faz necessário entender o porquê em trabalhos futuros. A qualidade da apresentação é identificada como uma dimensão adicional trazendo mais benefícios a respeito da clareza e organização com que os dados são exibidos e por isso sendo menos tratada nos estudos.

Não foi identificada uma relação direta entre os métodos de pré-processamento e as DQD propostas por Cai e Zhu (2015), a ausência de uma conexão explícita destaca a necessidade de uma investigação mais detalhada. Outro ponto a ser considerado é que a rápida evolução das tecnologias de *Big Data*, ML e DL podem exigir revisões periódicas para manter atualizadas as práticas relacionadas à qualidade de dados.

7 CONSIDERAÇÕES FINAIS

A qualidade dos dados é fundamental para obtenção dos melhores resultados em aplicações que fazem uso de tecnologias como *Big Data*, *Machine Learning* e *Deep Learning*, principalmente quando são aplicadas no âmbito da saúde, onde vidas podem ser impactadas em casos de erro. Com a revisão foi possível compreender os benefícios apontados pelos autores no uso destas tecnologias, entre elas estão a utilização devido ao grau de complexidade na análise de grandes volumes de dados, à melhoria na precisão dos resultados, à redução no tempo de resposta e à tomada de decisão orientada por dados.

Esta revisão de escopo buscou determinar métodos e *frameworks* aplicados à qualidade de dados médicos e de saúde para estas tecnologias. Os resultados obtidos permitiram o mapeamento e a categorização dos principais métodos de pré-processamento utilizados, assim como os seus benefícios e problemas mitigados. A limpeza de dados foi o método mais utilizado e também sua aplicação associada a outros métodos.

O estudo identificou as principais DQD abordadas nos artigos, permitindo entender que as métricas mais utilizadas foram a usabilidade, confiabilidade e disponibilidade para avaliar a qualidade dos dados.

Por outro lado, percebe-se que não existe uma relação direta entre os métodos e as dimensões, representando uma brecha na literatura e oportunidade para o desenvolvimento de *frameworks* que possam orientar de maneira clara a aplicação de métodos para atender às dimensões propostas por Cai e Zhu (2015).

REFERÊNCIAS

- ABIDI, Mustufa Haider et al. Big data-based smart health monitoring system: using deep ensemble learning. **IEEE Access**, 2023.
- ALHARBE, Nawaf; RAKROUKI, Mohamed Ali; ALJOHANI, Abeer. A healthcare quality assessment model based on outlier detection algorithm. **Processes**, v. 10, n. 6, p. 1199, 2022.
- ALOTAIBI, Nojod M.; ABDULLAH, Manal A. Big data mining: A classification perspective. In: **Communication, Management and Information Technology**. CRC Press, 2016. p. 701-710.
- ALZU'BI, Shadi et al. Diabetes monitoring system in smart health cities based on big data intelligence. **Future Internet**, v. 15, n. 2, p. 85, 2023.
- AN, Ning et al. A data-driven method for BioBrick quality assessment. In: **2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)**. IEEE, 2021. p. 83-87.
- AN, Ying et al. High-risk prediction of cardiovascular diseases via attention-based deep neural networks. **IEEE/ACM transactions on computational biology and bioinformatics**, v. 18, n. 3, p. 1093-1105, 2021.
- ANANTH, Antony Dennis; PALANISAMY, Chenniappan. Extended and optimized deep convolutional neural network-based lung tumor identification in big data. **International Journal of Imaging Systems and Technology**, v. 32, n. 3, p. 918-934, 2022.
- ARKSEY, Hilary; O'MALLEY, Lisa. Scoping studies: towards a methodological framework. **International journal of social research methodology**, v. 8, n. 1, p. 19-32, 2005.
- BAI, Shan et al. Study on the relationship between lung cancer stromal cells and air cavity diffusion based on an image acquisition system. **Contrast Media & Molecular Imaging**, v. 2022, n. 1, p. 2492124, 2022.
- BAUMGÄRTNER, Georg L. et al. Metadata-independent classification of MRI sequences using convolutional neural networks: Successful application to prostate MRI. **European Journal of Radiology**, v. 166, p. 110964, 2023.
- BECKER, David; KING, Trish Dunn; MCMULLEN, Bill. Big data, big data quality problem. In: **2015 IEEE international conference on big data (big data)**. IEEE, 2015, p. 2644-2653.
- CAI, Li; ZHU, Yangyong. The challenges of data quality and data quality assessment in the big data era. **Data science journal**, v. 14, p. 2-2, 2015.
- CARRA, Giorgia et al. Data-driven ICU management: Using Big Data and algorithms to improve outcomes. **Journal of critical care**, v. 60, p. 300-304, 2020.
- CERÓN, Juan Carlos Angeles et al. Assessing YOLACT++ for real time and robust instance segmentation of medical instruments in endoscopic procedures. In: **2021 43rd Annual**

International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, 2021. p. 1824-1827.

CHAKRABORTY, Chiranjib et al. From machine learning to deep learning: Advances of the recent data-driven paradigm shift in medicine and healthcare. **Current Research in Biotechnology**, v. 7, p. 100164, 2024.

CHEN, Xiaoyu; CAI, Qiang; CHEN, Xiaoming. ChroSegNet: An Attention-Based Model for Chromosome Instance Segmentation. In: **HBDSS 2022; 2nd International Conference on Health Big Data and Smart Sports**. VDE, 2022. p. 1-5.

CHOI, Jin.-Il.; OH, Jai.-Woo. A study of the automatic authentication module to improve the consistency of outsourced laboratory test results. **Journal of System and Management Sciences**, v. 12, n. 2, p. 397-410, 2022.

CHOI, Rene Y. et al. Introduction to machine learning, neural networks, and deep learning. **Translational vision science & technology**, v. 9, n. 2, p. 14-14, 2020.

CHU, Kuo-Chung et al. Modeling of Clinical Mammography Recognition. In: **2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI)**. IEEE, 2020. p. 409-411.

CHUNG, Kyungyong; YOO, Hyun; CHOE, Do-Eun. Ambient context-based modeling for health risk assessment using deep neural network. **Journal of Ambient Intelligence and Humanized Computing**, v. 11, n. 4, p. 1387-1395, 2020.

COELHO, Tatiane Perusse et al. Comparação e análise do uso de revisão sistemática e revisão de escopo na área do cuidado ao paciente na Farmácia. **Research, Society and Development**, v. 10, n. 12, p. e08101219915-e08101219915, 2021.

CORDEIRO, Luciana; SOARES, Cassia Baldini. Revisão de escopo: potencialidades para a síntese de metodologias utilizadas em pesquisa primária qualitativa. **Boletim do Instituto de Saúde - BIS**, v. 20, n. 2, p. 37-43, 2019.

DARGAN, Shaveta. A review on machine learning. *International Journal of Science and Research Archive*, 2023.

DARGAN, Shaveta et al. A survey of deep learning and its applications: a new paradigm to machine learning. **Archives of Computational Methods in Engineering**, v. 27, p. 1071-1092, 2020.

DASH, Sabyasachi et al. Big data in healthcare: management, analysis and future prospects. **Journal of big data**, v. 6, n. 1, p. 1-25, 2019.

DASU, Tamraparni. Data glitches: Monsters in your data. In: **Handbook of data quality: Research and practice**. Berlin, Heidelberg: Springer Berlin Heidelberg, p. 163-178, 2013.

DHAND, Geetika et al. Deep enriched salp swarm optimization based bidirectional-long short term memory model for healthcare monitoring system in big data. **Informatics in Medicine Unlocked**, v. 32, p. 101010, 2022.

EL RAHMAN, Sahar A.; ALLUHAIDAN, Ala Saleh; MARZOUK, Radwa. Breast Cancer Classification Based on DNA Microarray Analysis. *IEEE Access*, 2023.

FENG, Zhiwei et al. A preliminary study on projection denoising for low-dose CT imaging using modified dual-domain U-net. In: **2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD)**. IEEE, 2020. p. 223-226.

GHEISARI, Mehdi; WANG, Guojun; BHUIYAN, Md Zakirul Alam. A survey on deep learning in big data. In: **2017 IEEE international conference on computational science and engineering (CSE) and IEEE international conference on embedded and ubiquitous computing (EUC)**. IEEE, p. 173-180, 2017.

GU, Dongxiao et al. Visualizing the knowledge structure and evolution of big data research in healthcare informatics. *International journal of medical informatics*, v. 98, p. 22-32, 2017.

GUDIVADA, Venkat; APON, Amy; DING, Junhua. Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. *International Journal on Advances in Software*, v. 10, n. 1, p. 1-20, 2017.

GUPTA, Deepak; RANI, Rinkle. A study of big data evolution and research challenges. *Journal of information science*, v. 45, n. 3, p. 322-340, 2018.

HAAFZA, Laraib Aslam et al. Big data covid-19 systematic literature review: Pandemic crisis. *Electronics*, v. 10, n. 24, p. 3125, 2021.

HABEHH, Hafsa; GOHEL, Suril. Machine learning in healthcare. *Current genomics*, v. 22, n. 4, p. 291, 2021.

HOU, Yanrong et al. An ECG denoising method based on adversarial denoising convolutional neural network. *Biomedical Signal Processing and Control*, v. 84, p. 104964, 2023.

HSU, Ming-Hsia et al. Analyzing groups of inpatients' healthcare needs to improve service quality and sustainability. *Sustainability*, v. 13, n. 21, p. 11909, 2021.

INASTRILLA, Carlos Rafael Araujo. Big Data in Health Information Systems. In: **Seminars in Medical Writing and Education**. 2022. p. 1-6.

JAYAKARTHIK, R. et al. Fall Detection Scheme based on Deep Learning Model for High-Quality Life. In: **2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC)**. IEEE, 2022. p. 1582-1588.

JOHNSON, Justin M.; KHOSHGOFTAAR, Taghi M. Data-centric ai for healthcare fraud detection. *SN Computer Science*, v. 4, n. 4, p. 389, 2023.

JUDDOO, Suraj; GEORGE, Carlisle. A qualitative assessment of machine learning support for detecting data completeness and accuracy issues to improve data analytics in big data for the healthcare industry. In: **2020 3rd International Conference on Emerging Trends in Electrical, Electronic and Communications Engineering (ELECOM)**. IEEE, 2020. p. 58-66.

JUNG, Hoill; CHUNG, Kyungyong. Social mining-based clustering process for big-data integration. **Journal of Ambient Intelligence and Humanized Computing**, v. 12, n. 1, p. 589-600, 2021.

KIM, Dohyeong et al. Predicting asthma attacks: effects of indoor PM concentrations on peak expiratory flow rates of asthmatic children. **IEEE Access**, v. 8, p. 8791-8797, 2020.

KIM, Jungyoon; LIM, Jihye. A deep neural network-based method for prediction of dementia using big data. **International Journal of Environmental Research and Public Health**, v. 18, n. 10, p. 5386, 2021.

KIM, Mingyu et al. Deep learning in medical imaging. **Neurospine**, v. 16, n. 4, p. 657-668, 2019.

KIROLA, Madhu; MEMORIA, Minakshi; DUMKA, Ankur. Optimized U-Net convolutional neural network based breast cancer prediction for accuracy increment in big data. **Concurrency and Computation: Practice and Experience**, v. 35, n. 9, p. e7652, 2023.

KODIPALLI, Ashwini et al. A novel variant of deep convolutional neural network for classification of ovarian tumors using CT images. **Computers and Electrical Engineering**, v. 109, p. 108758, 2023.

KONG, Xiangyong et al. Disease-specific data processing: An intelligent digital platform for diabetes based on model prediction and data analysis utilizing big data technology. **Frontiers in Public Health**, v. 10, p. 1053269, 2022.

KRISHNAN, Krish. Data warehousing in the age of big data. **Newnes**, 2013.

KUANG, Hailan et al. Remote photoplethysmography signals enhancement based on generative adversarial networks. In: **2023 IEEE 3rd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA)**. IEEE, 2023. p. 792-796.

KUMAR, Akhilesh et al. Deep learning Based Patient-Friendly Clinical Expert Recommendation Framework. In: **2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM)**. IEEE, 2022. p. 736-741.

KUMAR, B. An encyclopedic overview of 'big data' analytics. **International Journal of Applied Engineering Research**, v. 10, n. 3, p. 5681-5705, 2015.

KUMAR, M. Jogendra et al. Examination Of Diabetes Mellitus For Ahead Of Schedule Expectation Utilizing Ideal Highlights Determination. In: **2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)**. IEEE, 2021. p. 1-6.

KURAL, Kamil Can et al. Using machine learning to improve anaphylaxis case identification in medical claims data. **JAMIA open**, v. 6, n. 4, p. ooad090, 2023.

LECUN, Yann; BENGIO, Yoshua; HINTON, Geoffrey. Deep learning. **nature**, v. 521, n. 7553, p. 436-444, 2015.

LEE, Chi Tak et al. The Precision in Psychiatry (PIP) study: Testing an internet-based methodology for accelerating research in treatment prediction and personalisation. **BMC psychiatry**, v. 23, n. 1, p. 25, 2023.

LEE, Kyu Hee et al. Concept and proof of the lifelog bigdata platform for digital healthcare and precision medicine on the cloud. *Yonsei medical journal*, v. 63, n. Suppl, p. S84, 2022.

LEI, Gaoyi et al. Comparative Analysis of Pre-process Pipelines For Automatic Retinal Vessel Segmentation. In: **2020 39th Chinese Control Conference (CCC)**. IEEE, 2020. p. 3216-3220.

LEVAC, Danielle; COLQUHOUN, Heather; O'BRIEN, Kelly K. Scoping studies: advancing the methodology. **Implementation science**, v. 5, p. 1-9, 2010.

LI, Jiahui et al. Measuring Sleep Stages and Quality Based on K-Means and Random Forest Algorithms. In: **2023 4th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT)**. IEEE, 2023. p. 484-488.

LI, Meng-Yi et al. Application of U-Net with Global Convolution Network Module in Computer-Aided Tongue Diagnosis. **Journal of healthcare engineering**, v. 2021, n. 1, p. 5853128, 2021.

LI, Tianhao et al. Study on the related factors of post-herpetic neuralgia in hospitalized patients with herpes zoster in Sichuan Hospital of Traditional Chinese Medicine based on big data analysis. **Dermatologic Therapy**, v. 33, n. 6, p. e14410, 2020.

LI, Wenrui et al. Multilevel stochastic optimization for imputation in massive medical data records. **IEEE Transactions on Big Data**, v. 10, n. 2, p. 122-131, 2023.

LIANG, NaNa. MRI image reconstruction based on artificial intelligence. In: **Journal of Physics: Conference Series**. IOP Publishing, 2021. p. 022077.

LIU, Junzhe et al. A diabetes prediction system based on federated learning. In: **2022 International Conference on Big Data, Information and Computer Network (BDICN)**. IEEE, 2022. p. 486-491.

LIU, Zheng et al. Heterogeneous similarity graph neural network on electronic health records. In: **2020 IEEE international conference on big data (big data)**. IEEE, 2020. p. 1196-1205.

LUO, Jake et al. Big data application in biomedical research and health care: a literature review. **Biomedical informatics insights**, v. 8, p. BII. S31559, 2016.

MAHANTI, Rupa. Data quality: dimensions, measurement, strategy, management, and governance. **Quality Press**, 2019.

MISHRA, Sushruta et al. Performance evaluation of a proposed machine learning model for chronic disease datasets using an integrated attribute evaluator and an improved decision tree classifier. **Applied Sciences**, v. 10, n. 22, p. 8137, 2020.

MOHAN, Maya et al. Secure Image Retrieval of Poor Quality Images by Combining LE-GAN, Arnold Mapping and Logistic Mapping. In: **2023 4th International Conference for Emerging Technology (INCET)**. IEEE, 2023. p. 1-6.

MORENO-SANCHEZ, Pedro A. Features importance to improve interpretability of chronic kidney disease early diagnosis. In: **2020 IEEE International Conference on Big Data (Big Data)**. IEEE, 2020. p. 3786-3792.

MÜFTÜOĞLU, Zümrüt; KIZRAK, M. Ayyüce; YILDLNM, Tülay. Differential privacy practice on diagnosis of COVID-19 radiology imaging using EfficientNet. In: **2020 international conference on innovations in intelligent systems and applications (inista)**. IEEE, 2020. p. 1-6.

MUMTAZ, Syeda Leena et al. Improvement of Dialysis Dosing Using Big Data Analytics. **Healthcare Informatics Research**, v. 29, n. 2, p. 174-185, 2023.

MUNAPPY, Aiswarya Raj et al. Data management for production quality deep learning models: Challenges and solutions. **Journal of Systems and Software**, v. 191, p. 111359, 2022.

MUNN, Zachary et al. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. **BMC medical research methodology**, v. 18, p. 1-7, 2018.

NICOLUCCI, Antonio et al. Prediction of complications of type 2 Diabetes: A Machine learning approach. **Diabetes Research and Clinical Practice**, v. 190, p. 110013, 2022.

NITHYA, T. et al. A comprehensive survey of machine learning: Advancements, applications, and challenges. In: **2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)**. IEEE, p. 354-361, 2023.

NNEJI, Grace Ugochi et al. A Dual Weighted Shared Capsule Network for Diabetic Retinopathy Fundus Classification. In: **2021 International Conference on High Performance Big Data and Intelligent Systems (HPBD&IS)**. IEEE, 2021. p. 297-302.

OH, J. W.; YOO, Og-Woo; KANG, Jin-Kyu. Automatic control system for quality assurance of clinical pathology examination: a machine learning approach. **Journal of System and Management Sciences**, v. 12, n. 1, p. 27-43, 2022.

OKTARINA, Eka Sari; SETIAWAN, Noor Akhmad; ARDIYANTO, Igi. A New Tele-Healthcare System of Heart Sound Signal Classification based on Gated Recurrent Unit. In: **2021 International Conference on Artificial Intelligence and Big Data Analytics**. IEEE, 2021. p. 1-6.

PAGE, Matthew J. et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. **bmj**, v. 372, 2021.

PAN, Yongsheng et al. Spatially-constrained fisher representation for brain disease identification with incomplete multi-modal neuroimages. **IEEE transactions on medical imaging**, v. 39, n. 9, p. 2965-2975, 2020.

PANDA, Sudhir et al. Health insurance cost prediction using regression models. In: **2022 International conference on machine learning, big data, cloud and parallel computing (COM-IT-CON)**. IEEE, 2022. p. 168-173.

PANICACCI, Silvia et al. Trading-off machine learning algorithms towards data-driven administrative-socio-economic population health management. **Computers**, v. 10, n. 1, p. 4, 2021.

PATI, Sarthak et al. Federated learning enables big data for rare cancer boundary detection. **Nature communications**, v. 13, n. 1, p. 7346, 2022

PRASINOS, Marios et al. A modelling framework for evidence-based public health policy making. **IEEE Journal of Biomedical and Health Informatics**, v. 26, n. 5, p. 2388-2399, 2022.

PRAVEEN, S. Phani et al. A robust framework for handling health care information based on machine learning and big data engineering techniques. **International Journal of Healthcare Management**, p. 1-18, 2022.

QIAN, Kun et al. Computer audition for fighting the SARS-CoV-2 corona crisis—Introducing the multitask speech corpus for COVID-19. **IEEE Internet of Things Journal**, v. 8, n. 21, p. 16035-16046, 2021.

QIN, Baoling. Research on Design of Fog Computing Optimization Model for Medical Big Data. In: **Computer Methods in Medicine and Health Care**. IOS Press, 2021. p. 93-102.

QIN, Chunli et al. Computer-aided detection in chest radiography based on artificial intelligence: a survey. **Biomedical engineering online**, v. 17, p. 1-23, 2018.

QU, Xiaolei et al. A deep learning-based automatic first-arrival picking method for ultrasound sound-speed tomography. **IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control**, v. 68, n. 8, p. 2675-2686, 2021.

RAGHUPATHI, Wullianallur; RAGHUPATHI, Viju. Big data analytics in healthcare: promise and potential. **Health information science and systems**, v. 2, p. 1-10, 2014.

RAGUPATHI, T.; GOVINDARAJAN, M.; PRIYARADHIKADEVI, T. Class Imbalance Handling with Deep Learning Enabled IoT Healthcare Diagnosis Model. **Intelligent Automation & Soft Computing**, v. 34, n. 2, 2022.

RAJESHKUMAR, K.; DHANASEKARAN, S.; VASUDEVAN, V. Efficient and secure medical big data management system using optimal map-reduce framework and deep learning. **Multimedia Tools and Applications**, v. 83, n. 16, p. 47111-47138, 2024.

RAJPURKAR, Pranav et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. **ArXiv abs/1711**, v. 5225, 2017.

RAO, Saka Uma Maheswara; RAO, K. Venkata; PVGD, Prasad Reddy. Medical Big Data Analysis using Binary Moth-Flame with Whale Optimization Approach. **International Journal of Advanced Computer Science and Applications**, v. 13, n. 8, 2022a.

RAO, Saka Uma Maheswara; RAO, K. Venkata; PVGD, Prasad Reddy. Medical Big Data Analysis using LSTM based Co-Learning Model with Whale Optimization Approach. **International Journal of Intelligent Engineering & Systems**, v. 15, n. 4, 2022b.

REHMAN, Amjad et al. Smart health analysis system using regression analysis with iterative hashing for IoT communication networks. **Computers and Electrical Engineering**, v. 104, p. 108456, 2022.

RIDZUAN, Fakhitah; ZAINON, Wan Mohd Nazmee Wan. A Review on Data Quality Dimensions for Big Data. **Procedia Computer Science**, v. 234, p. 341-348, 2024.

ROJAS, Esperanza Manrique et al. Contributions of machine learning in the health area as support in the diagnosis and care of chronic diseases. In: **Innovation in Medicine and Healthcare Systems, and Multimedia: Proceedings of KES-InMed-19 and KES-IIMSS-19 Conferences**. Springer Singapore, p. 261-269, 2019.

RUI, Xinkai et al. An Ontology-Driven Structured Processing Method for History of Present Illness of Discharge Summary. In: **2021 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCoM/CyberSciTech)**. IEEE, 2021. p. 419-427.

SAMARAS, Athanasios et al. Artificial intelligence-based mining of electronic health record data to accelerate the digital transformation of the national cardiovascular ecosystem: design protocol of the CardioMining study. **BMJ open**, v. 13, n. 4, p. e068698, 2023.

SARRAY, Noor; INAN, Timur. Prediction of tumor in mammogram images using data mining models. In: **2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)**. IEEE, 2022. p. 01-06.

SHRESTHA, Pratik et al. Early stage detection of scoliosis using machine learning algorithms. In: **2021 International Conference on Forensics, Analytics, Big Data, Security (FABS)**. IEEE, 2021. p. 1-4.

SHRIHARI, M. R. et al. Enhanced efficient and security in big data using TDES and machine learning technique. In: **2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)**. IEEE, 2022. p. 1-6.

SIVARAJAH, Uthayasankar et al. Critical analysis of Big Data challenges and analytical methods. **Journal of business research**, v. 70, p. 263-286, 2017.

SONI, Atul et al. Evaluating the Impact of Data Quality on Machine Learning Model Performance. **Journal of Nonlinear Analysis and Optimization**, v. 14, n. 01, 2023.

STEPHANIE, Veronika et al. Trustworthy privacy-preserving hierarchical ensemble and federated learning in healthcare 4.0 with blockchain. **IEEE Transactions on Industrial Informatics**, v. 19, n. 7, p. 7936-7945, 2022.

SUN, Hanxi et al. A deep learning approach to private data sharing of medical images using conditional generative adversarial networks (GANs). **Plos One**, v. 18, n. 7, p. e0280316, 2023.

SURANTHA, Nico; LESMANA, Tri Fennia; ISA, Sani Muhamad. Sleep stage classification using extreme learning machine and particle swarm optimization for healthcare big data. **Journal of Big Data**, v. 8, n. 1, p. 14, 2021.

SYED, Khajamoinuddin et al. Integrated natural language processing and machine learning models for standardizing radiotherapy structure names. In: **Healthcare**. MDPI, 2020. p. 120.

TAKURA, Tomoyuki; HIRANO GOTO, Keiko; HONDA, Asao. Development of a predictive model for integrated medical and long-term care resource consumption based on health behaviour: application of healthcare big data of patients with circulatory diseases. **BMC medicine**, v. 19, p. 1-16, 2021.

TAURION, Cezar. **Big data**. Brasport, 2013.

TAYLOR, Petroc. Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2023, with forecasts from 2024 to 2028. **Statista**, 2024. Disponível em: <https://www.statista.com/statistics/871513/worldwide-data-created/> Acesso em: 24 nov. 2024.

TAYLOR-SAKYI, Kevin. Big data: Understanding big data. **arXiv preprint arXiv:1601.04602**, 2016.

THIEME, Anja; BELGRAVE, Danielle; DOHERTY, Gavin. Machine learning in mental health: A systematic review of the HCI literature to support the development of effective and implementable ML systems. **ACM Transactions on Computer-Human Interaction (TOCHI)**, v. 27, n. 5, p. 1-53, 2020.

TOMAZ SANTOS, Livia C.; BUBLITZ, Frederico M. Data quality and Big Data in the health industry: a scoping review protocol. **medRxiv**, p. 2024.10. 18.24315741, 2024.

TRICCO, Andrea C. et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. **Annals of internal medicine**, v. 169, n. 7, p. 467-473, 2018

TRIPATHI, Rajan Prasad et al. Timely prediction of diabetes by means of machine learning practices. **Augmented Human Research**, v. 8, n. 1, p. 1, 2023.

TSOU, Ching-Shih et al. Quality prediction through machine learning for the inspection and manufacturing process of blood glucose test strips. **Cogent Engineering**, v. 9, n. 1, p. 2083475, 2022.

TUMMALA, Jagadeesh et al. 3D Deep Convolutional Neural Network for Detection of Anomalous Rigid and Affine Registrations in Big Data Brain MRI. In: **2021 IEEE Bombay Section Signature Conference (IBSSC)**. IEEE, 2021. p. 1-6.

VALDÉS, Julio et al. Cough Classification with Deep Derived Features using Audio Spectrogram Transformer. In: **2022 IEEE International Conference on Big Data (Big Data)**. IEEE, 2022. p. 1729-1739.

WANG, Jingran et al. Overview of data quality: Examining the dimensions, antecedents, and impacts of data quality. **Journal of the Knowledge Economy**, v. 15, n. 1, p. 1159-1178, 2024.

WANG, Shuqiang et al. Diabetic retinopathy diagnosis using multichannel generative adversarial network with semisupervision. **IEEE Transactions on Automation Science and Engineering**, v. 18, n. 2, p. 574-585, 2021.

WEVODAU, Z. et al. The Temple University Digital Pathology Corpus: The Breast Tissue Subset. In: **2021 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)**. IEEE, 2021. p. 1-4.

WU, Menglin et al. Noise reduction for SD-OCT using a structure-preserving domain transfer approach. **IEEE Journal of Biomedical and Health Informatics**, v. 25, n. 9, p. 3460-3472, 2021.

XU, Chenchu et al. Applying cross-modality data processing for infarction learning in medical internet of things. **IEEE Internet of Things Journal**, v. 8, n. 23, p. 16902-16910, 2021.

YAN, Da et al. Improving brain dysfunction prediction by gan: A functional-connectivity generator approach. In: **2021 IEEE International Conference on Big Data (Big Data)**. IEEE, 2021. p. 1514-1522.

YANG, Qiming et al. Mining domain knowledge: improved framework towards automatically standardizing anatomical structure nomenclature in radiotherapy. **IEEE Access**, v. 8, p. 105286-105300, 2020.

YANG, Shuojue et al. Learning COVID-19 pneumonia lesion segmentation from imperfect annotations via divergence-aware selective training. **IEEE Journal of Biomedical and Health Informatics**, v. 26, n. 8, p. 3673-3684, 2022.

ZEBARI, Dilovan Asaad et al. Automated Detection of Covid-19 from X-ray Using SVM. In: **2022 4th International Conference on Advanced Science and Engineering (ICOASE)**. IEEE, 2022. p. 130-135.

ZENG, Yan; LIANG, Wei. Research on Key Technologies of Multi-modal Heterogeneous Medical Big Data Based on Deep Learning. In: **ISCTT 2022; 7th International Conference on Information Science, Computer Technology and Transportation**. VDE, 2022. p. 1-6.

ZHANG, Baoshe et al. A systematic quality assurance framework for the upgrade of radiation oncology information systems. **Physica Medica**, v. 69, p. 28-35, 2020.

ZHANG, Mo et al. Automated semantic segmentation of red blood cells for sickle cell disease. **IEEE journal of biomedical and health informatics**, v. 24, n. 11, p. 3095-3102, 2020.

ZHANG, Weiguo; CUI, Linfang. Detection algorithm of pulmonary nodules based on deep learning. In: **2021 2nd International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)**. IEEE, 2021. p. 185-188.

ZHANG, Yuanyu; LI, Yunqi; HE, Dongzhi. U-shaped network with multi-scale feature extraction for lesion region segmentation in gastric cancer images. In: **2023 IEEE 3rd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA)**. IEEE, 2023. p. 234-238.

ZHAO, Shangping et al. Development and application of an intensive care medical data set for deep learning. In: **2020 IEEE International Conference on Big Data (Big Data)**. IEEE, 2020. p. 3369-3373.

ZHOU, Qingguo et al. Hierarchical Hybrid Networks for Automatic Pulmonary Blood Vessel Segmentation in Computed Tomography Images. **IEEE/ACM Transactions on Computational Biology and Bioinformatics**, 2023.

APÊNDICE A - PROTOCOLO PUBLICADO

TOMAZ SANTOS, Livia C.; BUBLITZ, Frederico M. Data quality and Big Data in the health industry: a scoping review protocol. *medRxiv*, p. 2024.10. 18.24315741, 2024.

medRxiv preprint doi: <https://doi.org/10.1101/2024.10.18.24315741>; this version posted October 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Data quality and Big Data in the health industry: a scoping review protocol

Livia C. T Santos^{1,*}, Frederico M. Bublitz¹

¹State University of Paraíba – Center for Strategic Technologies in Health (NUTES), Campina Grande, PB, Brazil.

*Corresponding Author: livia.caroline.tomaz.santos@aluno.uepb.edu.br

ABSTRACT

Introduction: Big Data is characterized by the large volume of data, the variety of types and formats, the speed with which they are generated, and the veracity and value that can be extracted from the data. However, the result obtained with this technology will depend on the quality of the information obtained from the data. Big Data has great potential in healthcare and can be used to advance diagnosis, treatment, and healthcare management. Health data is highly vulnerable due to its sensitive nature, as it contains personal and confidential information. If exposed or compromised, it could lead to privacy violations, inaccuracies, misuse, incorrect diagnoses, or misguided decision-making in patient care. It is important to prioritize confidentiality, adhere to regulatory compliance, and maintain data integrity; for that, it is essential to use efficient methods to obtain quality data and make them able to reach the proposed objective.

Objective: In this context, the scoping review protocol aims to identify and map existing strategies, methods, or models that improve the quality of medical and health data in Big Data environments. This review explores the methods to support the effective use of Big Data in healthcare while addressing the challenges to maintain data integrity and ensure safe decision-making.

Methods and analysis: This scoping review will be conducted based on the six-step process outlined in the framework proposed by Levac et al. in "Scoping Studies: Advancing the methodology" and will be reported following the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) checklist. The research team will use Data Quality, Big Data, and Health terms to search for primary studies in the Scopus Document Search, IEEE Xplore Digital Library, and ACM Digital Library databases.

Keywords: Scoping Review; Big Data; Health Data; Data Quality.

INTRODUCTION

Data generation has significantly increased in today's digital world due to the diverse methods available for acquiring, managing, and processing information [1]. This has led businesses and organizations to face the challenge of handling large amounts of data from various sources while extracting valuable insights from it [2]. Taurion [3] defines Big Data as an equation combining volume, variety, velocity, and veracity, with all these factors adding value. The concept of Big Data is characterized by the 5 V's [4], which capture its essential features: (i) Volume refers to the large amount of data generated and obtained; (ii) Variety refers to the data types and formats available; (iii) Velocity refers to the rapid rate at which data is produced, transmitted, stored, and processed, emphasizing the need for real-time or

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

near-real-time analysis; (iv) Veracity addresses the consistency and trustworthiness of the data; (v) Value refers to the insights and potential advantages derived from the data.

Big Data allows for more data-driven decision-making and innovation across numerous sectors. These innovations include using Large Language Models (LLMs) that use machine learning and deep learning to analyze vast amounts of text data and generate human-like language, detect patterns, and provide insights [5]. LLM can be applied to medicine, assisting with drafting medical documentation and diagnosis and improving communication between healthcare providers and patients [6].

In healthcare, Big data plays a critical role by supporting the analysis of large datasets from electronic health records, medical imaging, and wearable devices to help professionals make decisions and improve patient diagnosis and treatment. However, ensuring the reliability of health data remains a challenge, as the quality of these vast datasets directly impacts the effectiveness of big data-driven healthcare solutions in life-or-death decisions [7]. Some benefits of using this technology in healthcare include disease prediction [8,9] and enhancing personalized treatment for individual patients through Precision Medicine [10].

The size and speed of data generation also present challenges, particularly regarding data quality, integration, and privacy concerns. Cai and Zhu [11] highlight several challenges related to big data quality, including integrating diverse and complex data types that complicate the overall data management process; as the volume of data overgrows, evaluating its quality becomes increasingly challenging; furthermore, the fast-changing nature of data requires agile processing technologies to keep up with constant changes; another critical issue is the lack of unified standards for assessing data quality.

This article presents a scoping review protocol to address these challenges and improve the understanding of data quality and Big Data in the health industry. The protocol aims to systematically explore the existing literature on data quality and enhance data quality management strategies to promote the better use of big data in healthcare. This review will allow us to map the main problems or concerns with methods that mitigate those problems. Additionally, understanding the gaps in data quality research within the health industry is critical for guiding future research.

Previous systematic reviews

During the search rounds to define the terms used in the search string, no systematic reviews were identified whose emphasis is understanding and listing the main models, frameworks, or criteria used in data quality for Big Data, whose focus is health systems.

Previous reviews have focused on identifying this technology's benefits, issues, and possible applications in health [12-15].

METHODS AND ANALYSIS

The scoping review will be developed following the methodological framework proposed by Levac et al. [16] in "Scoping studies: Advancing the methodology", which provides recommendations to clarify and

improve the scoping review methodology presented by Arksey and O'Malley [17]. The framework has six stages:

1. Identifying the research question.
2. Identifying relevant studies.
3. Study selection.
4. Charting the data.
5. Collating, summarising, and reporting the results.
6. Consultation (optional stage).

The PRISMA-ScR [18] checklist will also be used to report the results.

1. Identifying the research question

This research aims to determine criteria, processes, methods, or models applied to medical and health data quality for Big Data. From the proposed objective, this research seeks to answer the following questions:

- What are the main problems or concerns about the quality of Big Data in the health industry? As an initial step in the study of the use of big data in health, it is necessary to understand the difficulties encountered in using this technology in healthcare.
- What methods are used for data quality? This question aimed to identify the methods used in papers and their possible benefits.
- What problems with data quality does this method mitigate? Identifying the mitigated issues will allow us to categorize the challenges encountered using this technology.
- What gaps exist in the health industry's data quality research for Big Data? Understanding the gaps will provide information for future work in this study area.

2. Identifying relevant studies

This review will be conducted in the Scopus Document Search¹, IEEE XploreDigital Library², and ACM Digital Library³ databases, which are multidisciplinary databases with extensive work related to this review's scope.

The team will conduct a series of searches for primary studies relevant to the research with different terms associated with data quality, big data, and health.

As a result of rounds of search and evaluation of the relevance of the papers obtained, the final search string was defined using the terms (("big data") AND ("clean*" OR "quality" OR "dirty") AND ("criteria" OR "process" OR "method" OR "model") AND ("health*" OR "medic*")) in databases.

3. Study selection

¹ <https://www.scopus.com/home.uri>

² <https://ieeexplore.ieee.org/Xplore/home.jsp>

³ <https://dl.acm.org/>

This review aims to identify different models, frameworks, or criteria for improving the quality of medical and health data in Big Data. The inclusion (INC) and exclusion (EXC) criteria in Table 1 will guide the search and review of the papers obtained and determine the study's eligibility.

Table 1: Inclusion and Exclusion Criteria

Inclusion Criteria	Exclusion Criteria
INC.01 - Domain: The article must address criteria, processes, methods, or models applied to the quality of health data for Big Data. INC.02 - Language: The article must be written in English. INC.03 - Date: There are no date restrictions. INC.04 - Study design: The article presents a primary study. INC.05 - Availability: The complete article is available⁴.	EXC.01 - Duplicity: The same article is found in another database. In this scenario, duplicates will be removed. EXC.02 - Does not meet all the inclusion criteria.

With all the articles obtained from the searches in the databases and the eligibility criteria defined, the team will carry out the following steps:

1. Remove duplicate articles, which will be automatically identified after importing search results into Rayyan⁵, software for managing systematic reviews.
2. Reading the title and abstract and removing those studies that do not meet the inclusion criteria.
3. Perform a full-text review, excluding studies that do not meet the criteria and explaining the reasons for exclusion.

In case of conflict in the steps, the team member will discuss with the others to find a solution. The results obtained after following all the steps will complete the PRISMA flow diagram, which shows the number of articles identified, duplicated, screened, eligible, and included in the review.

4. Charting the data

After selecting the studies, data will be extracted and charted from the articles that meet the inclusion criteria. Two researchers will collect the information separately using a data-charting form and continually update it to expand as new important data is identified [16] and later merge it. The form will include:

⁴ Articles must be available in the databases linked to the Portal of Journals of CAPES (Coordination for the Improvement of Higher Education Personnel).

⁵ <https://www.rayyan.ai/>

- Article title
- Author(s)
- Country
- Year
- Publisher
- Study objectives
- Methodologies
- Results
- Data quality methods
- Characteristics of the techniques
- Data quality problems mitigated
- Study strengths and limitations
- Team conclusion about the article

The data-charting form will identify the main problems related to Big Data quality in the health industry by summarizing results and the specific quality issues addressed in each article. The methods used for data quality will be captured, allowing for a comparative analysis of their effectiveness and benefits. Additionally, by detailing the characteristics of these techniques and the problems mitigated, the review will categorize challenges encountered in utilizing Big Data in healthcare. Lastly, insights on study strengths and limitations will reveal gaps in current research, guiding future work in this area.

5. Collating, summarising and reporting the results

As defined in the methodological framework by Levac et al. [16], this stage contains three steps and will allow researchers to obtain meaningful information from the data collected.

- Analysis: The team will provide a qualitative and quantitative analysis of the extracted information. Qualitative can give a better overview of the techniques provided and open up new study opportunities. In the quantitative analysis, the classification of techniques will be carried out according to their characteristics and the mitigated problems;
- Reporting the results: The team will present the results using tables with findings and gaps;
- Findings: The aim is to discuss these results and possible applications in future research.

DISCUSSION

This scoping review protocol will synthesize the existing literature and guide future reviews to answer the questions raised. It aims to identify the methods used to improve health and medical data quality for Big Data, determine the most used, and address the problems that mitigate them. The protocol will also map the principal concepts and gaps associated with the topic. The results obtained will help future work in this area of study.

COMPETING INTERESTS

The authors declare no competing interests.

AUTHOR CONTRIBUTIONS

All team members contributed equally to the conceptualization, development, and writing of the draft of the manuscript.

FUNDING

The authors received no funding for this research.

ETHICS AND DISSEMINATION

No ethics approval is required for this scoping review protocol, as primary data will not be collected. The review will produce important outcomes in improving the quality of health data for Big Data. The results obtained from this scoping review will be published in a peer-reviewed journal.

REFERENCES

- [1] Gandhi, K. I., & Sreedhar, S. C. (2015). Survey on Big Data: Management and Challenges. *Int. J. Comput. Trends Technol*, 20, 33-36.
- [2] Medida, L. H. (2024). Addressing Challenges in Data Analytics: A Comprehensive Review and Proposed Solutions. *Critical Approaches to Data Engineering Systems and Analysis*, 16-33.
- [3] Taurion, C. (2013). *Big data*. Brasport.
- [4] Kumar, B. (2015). An encyclopedic overview of 'big data' analytics. *International Journal of Applied Engineering Research*, 10(3), 5681-5705.
- [5] Hadi, M. U., Al Tashi, Q., Shah, A., Qureshi, R., Muneer, A., Irfan, M., ... & Shah, M. (2024). Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints*.
- [6] Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature medicine*, 29(8), 1930-1940.
- [7] Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health information science and systems*, 2, 1-10.
- [8] AlZu'bi, S., Elbes, M., Mughaid, A., Bdair, N., Abualigah, L., Forestiero, A., & Zitar, R. A. (2023). Diabetes monitoring system in smart health cities based on big data intelligence. *Future Internet*, 15(2), 85.
- [9] Ananth, A. D., & Palanisamy, C. (2022). Extended and optimized deep convolutional neural network-based lung tumor identification in big data. *International Journal of Imaging Systems and Technology*, 32(3), 918-934.

- [10] Yang, X., Huang, K., Yang, D., Zhao, W., & Zhou, X. (2024). Biomedical big data technologies, applications, and challenges for precision medicine: A review. *Global Challenges*, 8(1), 2300163.
- [11] Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data science journal*, 14, 2-2.
- [12] Luo, J., Wu, M., Gopukumar, D., & Zhao, Y. (2016). Big data application in biomedical research and health care: a literature review. *Biomedical informatics insights*, 8, BII-S31559.
- [13] Gaitanou, P., Garoufallou, E., & Balatsoukas, P. (2014). The effectiveness of big data in health care: a systematic review. In *Metadata and Semantics Research: 8th Research Conference, MTSR 2014, Karlsruhe, Germany, November 27-29, 2014. Proceedings 8* (pp. 141-153). Springer International Publishing.
- [14] Raja, R., Mukherjee, I., & Sarkar, B. K. (2020). A systematic review of healthcare big data. *Scientific Programming*, 2020, 1-15.
- [15] Singh, R. K., Agrawal, S., Sahu, A., & Kazancoglu, Y. (2023). Strategic issues of big data analytics applications for managing health-care sector: a systematic literature review and future research agenda. *The TQM Journal*, 35(1), 262-291.
- [16] Levac, D., Colquhoun, H., & O'Brien, K. K. (2010). Scoping studies: advancing the methodology. *Implementation Science*, 5, 1-9.
- [17] Arksey, H., & O'Malley, L. (2005). Scoping studies: towards a methodological framework. *International journal of social research methodology*, 8(1), 19-32.
- [18] Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., ... & Straus, S. E. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Annals of internal medicine*, 169(7), 467-473.

ANEXO A - PRISMA ScR

Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) Checklist

SECTION	ITEM	PRISMA-ScR CHECKLIST ITEM	REPORTED ON PAGE #
TITLE			
Title	1	Identify the report as a scoping review.	1
ABSTRACT			
Structured summary	2	Provide a structured summary that includes (as applicable): background, objectives, eligibility criteria, sources of evidence, charting methods, results, and conclusions that relate to the review questions and objectives.	6,7
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known. Explain why the review questions/objectives lend themselves to a scoping review approach.	10-13
Objectives	4	Provide an explicit statement of the questions and objectives being addressed with reference to their key elements (e.g., population or participants, concepts, and context) or other relevant key elements used to conceptualize the review questions and/or objectives.	13
METHODS			
Protocol and registration	5	Indicate whether a review protocol exists; state if and where it can be accessed (e.g., a Web address); and if available, provide registration information, including the registration number.	22, 23
Eligibility criteria	6	Specify characteristics of the sources of evidence used as eligibility criteria (e.g., years considered, language, and publication status), and provide a rationale.	24, 25
Information sources*	7	Describe all information sources in the search (e.g., databases with dates of coverage and contact with authors to identify additional sources), as well as the date the most recent search was executed.	24
Search	8	Present the full electronic search strategy for at least 1 database, including any limits used, such that it could be repeated.	24
Selection of sources of evidence†	9	State the process for selecting sources of evidence (i.e., screening and eligibility) included in the scoping review.	24, 25
Data charting process‡	10	Describe the methods of charting data from the included sources of evidence (e.g., calibrated forms or forms that have been tested by the team before their use, and whether data charting was done independently or in duplicate) and any processes for obtaining and confirming data from investigators.	25, 26
Data items	11	List and define all variables for which data were sought and any assumptions and simplifications made.	26
Critical appraisal of individual sources of evidence§	12	If done, provide a rationale for conducting a critical appraisal of included sources of evidence; describe the methods used and how this information was used in any data synthesis (if appropriate).	N/A
Synthesis of results	13	Describe the methods of handling and summarizing the data that were charted.	26

SECTION	ITEM	PRISMA-ScR CHECKLIST ITEM	REPORTED ON PAGE #
RESULTS			
Selection of sources of evidence	14	Give numbers of sources of evidence screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally using a flow diagram.	27
Characteristics of sources of evidence	15	For each source of evidence, present characteristics for which data were charted and provide the citations.	28-38
Critical appraisal within sources of evidence	16	If done, present data on critical appraisal of included sources of evidence (see item 12).	N/A
Results of individual sources of evidence	17	For each included source of evidence, present the relevant data that were charted that relate to the review questions and objectives.	38-43
Synthesis of results	18	Summarize and/or present the charting results as they relate to the review questions and objectives.	38-44
DISCUSSION			
Summary of evidence	19	Summarize the main results (including an overview of concepts, themes, and types of evidence available), link to the review questions and objectives, and consider the relevance to key groups.	45
Limitations	20	Discuss the limitations of the scoping review process.	46
Conclusions	21	Provide a general interpretation of the results with respect to the review questions and objectives, as well as potential implications and/or next steps.	46
FUNDING			
Funding	22	Describe sources of funding for the included sources of evidence, as well as sources of funding for the scoping review. Describe the role of the funders of the scoping review.	N/A

JBI = Joanna Briggs Institute; PRISMA-ScR = Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews.

* Where *sources of evidence* (see second footnote) are compiled from, such as bibliographic databases, social media platforms, and Web sites.

† A more inclusive/heterogeneous term used to account for the different types of evidence or data sources (e.g., quantitative and/or qualitative research, expert opinion, and policy documents) that may be eligible in a scoping review as opposed to only studies. This is not to be confused with *information sources* (see first footnote).

‡ The frameworks by Arksey and O'Malley (6) and Levac and colleagues (7) and the JBI guidance (4, 5) refer to the process of data extraction in a scoping review as data charting.

§ The process of systematically examining research evidence to assess its validity, results, and relevance before using it to inform a decision. This term is used for items 12 and 19 instead of "risk of bias" (which is more applicable to systematic reviews of interventions) to include and acknowledge the various sources of evidence that may be used in a scoping review (e.g., quantitative and/or qualitative research, expert opinion, and policy document).

From: Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Ann Intern Med*. 2018;169:467–473. doi: [10.7326/M18-0850](https://doi.org/10.7326/M18-0850).