

Um Estudo sobre Melhorias na Qualidade dos Dados para Análises no Domínio da Saúde

Júlia Nakayama Silveira¹, Ronaldo dos Santos Mello¹

¹Universidade Federal de Santa Catarina (UFSC), INE, PPGCC, Florianópolis, SC

juanakayama97@gmail.com

r.mello@ufsc.br

Abstract. *The heterogeneity, fragmentation, and secondary use of clinical data make data quality a central requirement for advanced analytics and artificial intelligence applications in healthcare. This paper presents a study on data quality improvement strategies aimed at enabling more effective analyses in the Health domain. The study includes a systematic literature review to identify strategies, dimensions and metrics, technologies, challenges, and outcomes associated with data quality assessment and improvement, followed by an overview and comparative analysis of the selected works. The findings indicate convergence around the use of classical dimensions such as completeness and accuracy, as well as aggregated indicators to support comparable measurement. The results reveal a technological evolution from validation dashboards and frameworks toward machine learning-based solutions, real-time monitoring architectures, and emerging approaches using LLM and federated learning.*

Resumo. *A heterogeneidade, fragmentação e uso secundário de dados clínicos tornam a qualidade de dados um requisito central para análises avançadas e aplicações de inteligência artificial em saúde. Este artigo apresenta um estudo sobre melhorias na qualidade de dados visando análises mais efetivas sobre esses dados, considerando o domínio de Saúde. O estudo inclui uma revisão sistemática da literatura para identificar estratégias, dimensões/métricas, tecnologias, desafios e resultados associados à avaliação e melhoria da qualidade de dados do domínio, seguido de uma visão geral e uma análise comparativa dos trabalhos encontrados. Os resultados do estudo indicam convergência no uso de dimensões clássicas, como completude e acurácia, e de indicadores agregados para mensuração comparável. Observa-se também a evolução tecnológica de dashboards e frameworks de validação para soluções baseadas em aprendizado de máquina, monitoramento em tempo real e abordagens emergentes que utilizam LLM e aprendizado federado.*

1. Introdução

A gestão de *Big Data* exige o tratamento da heterogeneidade estrutural, semântica e de qualidade dos dados, pois a diversidade de formatos e fontes amplia riscos de inconsistência e compromete a confiabilidade analítica. Nesse contexto, governança e veracidade tornam-se pré-requisitos para análises avançadas e aplicações de inteligência artificial [Shahnawaz and Kumar 2025].

No domínio da *Saúde*, essa complexidade é intensificada pela integração de dados epidemiológicos, clínicos, demográficos e ambientais, frequentemente descentralizados e com diferentes níveis de qualidade. Pesquisas que aplicam aprendizado de máquina (AM) dependem dessa harmonização, conforme demonstrado por [Joshi et al. 2023], que evidencia o impacto do AM na detecção automatizada de irregularidades em fluxos de dados de saúde para apoiar a vigilância e a tomada de decisão crítica. A literatura reforça que a avaliação sistemática da qualidade é essencial para decisões seguras e melhor desempenho organizacional [Kaloyanova et al. 2021] [Nasir et al. 2022], evitando erros que podem comprometer tanto a gestão quanto o cuidado ao paciente [Nasir et al. 2022]. Além disso, a melhoria contínua dos dados é crucial para garantir validade em pesquisas e análises secundárias, exigindo validação e controle de vieses [Kaloyanova et al. 2021, Goldstein et al. 2022].

Diante disso, este trabalho apresenta um estudo sobre o controle de qualidade de dados no domínio de Saúde, incluindo uma análise comparativa das abordagens identificadas e sugestões de direcionamento futuro em termos de pesquisa. Este artigo está organizado conforme segue. A Seção 2 descreve a revisão sistemática da literatura realizada. A Seção 3 apresenta uma visão geral dos estudos selecionados. A Seção 4 mostra uma análise comparativa dos estudos selecionados e a Seção 5 traz as conclusões.

2. Revisão Sistemática da Literatura

Uma revisão sistemática da literatura (RSL) foi conduzida para identificar, sintetizar e analisar os principais estudos relacionados à avaliação e melhoria da qualidade dos dados na área da Saúde. Esta revisão seguiu o protocolo proposto por [Kitchenham 2004], amplamente utilizado em pesquisas na área de Ciência da Computação. Assim sendo, foram definidas as seguintes questões de pesquisa (QP):

- **QP1:** Quais estratégias tecnológicas de melhorias na qualidade dos dados têm sido utilizadas para apoiar análises da área da Saúde?
- **QP2:** Quais são as principais dimensões e métricas de qualidade consideradas?
- **QP3:** Quais ferramentas e tecnologias têm sido propostas para melhorar a qualidade dos dados na área da Saúde?
- **QP4:** Quais são os principais desafios técnicos e operacionais nas tecnologias voltadas à melhoria da qualidade dos dados da área da Saúde?
- **QP5:** Quais foram os resultados das melhorias aplicadas?

A busca por estudos foi realizada nas seguintes bases de dados científicas amplamente reconhecidas: *IEEE Xplore*, *ScienceDirect*, *Scopus*, *SpringerLink* e *Pub Med*. Na sequência, os seguintes critérios de inclusão foram definidos: (i) Estudos publicados entre 2020 e 2026; (ii) Estudos publicados em veículos bem qualificados; (iii) Estudos com foco em avaliar e melhorar a qualidade de dados; (iv) Estudos com a temática relacionada à área da saúde; (v) Estudos disponíveis na íntegra. A próxima etapa da revisão foi a definição de *strings* de busca, elaboradas a partir de termos-chave relacionados ao foco do estudo. A Tabela 1 apresenta, para cada uma das bases de dados, a *string* de busca utilizada e os resultados da busca.

Table 1. Estudos identificados e selecionados na RSL

Repositório	String de Busca	Resultado	Seleção
IEEE Xplore	(((((“Document Title”: data) AND (“Document Title”: quality) AND (“All Metadata”: healthcare OR “All Metadata”: “public health” OR “All Metadata”: “electronic health record” OR “All Metadata”: “health system”))))))	16	3
Science Direct	Year: 2020–2026. Title/Abstract/Keywords: healthcare OR “public health” OR “electronic health record” OR “health system”. Title: data AND quality.	14	1
Scopus	TITLE(data) AND TITLE(quality) AND (TITLE(healthcare) OR TITLE(“public health”) OR TITLE(“electronic health record”) OR TITLE(“health system”)) AND PUBYEAR > 2019 AND PUBYEAR < 2027 AND LIMIT-TO(SUBJAREA,“COMP”) AND (LIMIT-TO(DOCTYPE,“ar”) OR LIMIT-TO(DOCTYPE,“cp”)).	58	9
Springer	data AND quality AND (“public health” OR healthcare OR “health system” OR “health information system” OR his OR “electronic health record”).	40	3
PubMed	data[Title] AND quality[Title] AND (“public health”[Title] OR healthcare[Title] OR “health system”[Title] OR “health information system”[Title] OR his[Title] OR “electronic health record”[Title]).	141	6
Total		269	22

Conforme observado, 269 trabalhos foram encontrados. Após a leitura dos títulos e dos resumos, os seguintes critérios de exclusão foram aplicados: (i) Estudos duplicados; (ii) Estudos com foco em revisão do estado da arte; (iii) Estudos sem foco em qualidade de dados da área da Saúde; (iv) Estudos que não publicados em revistas ou conferências; (v) Estudos teóricos ou sem experimento prático; (vi) Estudos sem viés tecnológico; (vii) Estudos não disponíveis na íntegra. Após a aplicação destes critérios de exclusão permaneceram 22 estudos, que são detalhados na próxima seção.

3. Visão Geral dos Estudos Selecionados

Grande parte da literatura concentra-se na definição de *modelos e indicadores* para mensurar a qualidade de dados clínicos e de saúde. Exemplos incluem o *DQ4HEALTH*, com dimensões e regras de validação para dados no padrão *OMOP CDM* [Kim et al. 2021], além de *frameworks* aplicados à avaliação de fenótipos específicos, como *DCIS* [Berman et al. 2024], e à validação de *EHRs* (*Electronic Health Record* - Registro Eletrônico de Saúde) para modelos preditivos clínicos [Thuraisingam et al. 2021].

Metodologias estatísticas avançadas têm sido aplicadas para avaliar a plausibilidade de dados longitudinais [Wang et al. 2021] e estabelecer padrões de qualidade, como o *DataGauge* para risco cardiovascular [Diaz-Garelli et al. 2019]. Experiências em redes como a *CPCSSN* destacam a necessidade de monitoramento contínuo [Ehsani-Moghaddam et al. 2021], enquanto em hospitais, *métricas ponderadas* apoiam a geração de escores de qualidade para decisão [Monteiro et al. 2025]. Iniciativas como a *MIDAS* [Maity et al. 2024] e *etiquetas contextuais* para dados genômicos [Griffiths et al. 2024] ampliam o controle da qualidade para uso em IA.

O uso de *AM* tem aprimorado a detecção de anomalias e o tratamento de inconsistências em dados de saúde. O *framework FlaSH* permite identificar *outliers* em tempo real [Joshi et al. 2023], enquanto o *PyCaret* apresenta desempenho superior a métodos tradicionais da OMS em dados de HIV/AIDS [Olaniyan and Owoseni 2022]. Além disso, o *ML-DQA* atua como um sistema estruturado de garantia de qualidade antes do desenvolvimento de modelos [Sendak et al. 2022]. Estratégias de *AM* também têm sido empregadas para aprimorar acurácia e completude dos dados [Jarmakovica 2025], incluindo *algoritmos de imputação baseados em redes neurais não supervisionadas*, como

o *CLUSTIMP*, para tratamento de valores ausentes [Shobha and Savarimuthu 2021]. Tecnologias emergentes, como *Modelos de Linguagem de Grande Escala (LLMs)* no sistema *LLM-DQR*, possibilitam a geração automática de regras complexas de qualidade [Xie et al. 2025]. Além disso, destacam-se abordagens como *Untrainable Data Cleansing (UDC)* [Hall et al. 2021] e *AM Federado (AF)* para dados sensíveis ao contexto [Jeon et al. 2023].

A automação da integração de dados e a otimização de fluxos ETL são críticas para a interoperabilidade em larga escala. Estudos propõem *frameworks* de *Master Data Management (MDM)*, gestão de um subconjunto de dados consistente e uniforme de entidades de negócio que descrevem as atividades principais de uma empresa, integrados a modelagens BPMN, que alinham a qualidade dos dados aos processos de negócio hospitalares [Benkherourou and Bourouis 2025]. Além disso, *arquiteturas de monitoramento em tempo real com ELK Stack* têm sido adotadas para assegurar a integridade dos *logs* de ingestão [Kamal et al. 2025].

A aplicação de painéis interativos, como o *Data Quality Dashboard (DQD)* da OHDSI, têm apoiado a manutenção de armazéns de dados de pesquisa [Marteau et al. 2025]. Para notas clínicas semiestruturadas, *técnicas de mineração de texto e regex* viabilizam a integração automatizada de dados, como em estudos sobre epilepsia [Fernandes et al. 2022]. Por fim, o desenvolvimento de *registros ágeis* em centros de saúde comunitários demonstrou como superar desafios de infraestrutura para gerar dados padronizados durante crises sanitárias [Romero et al. 2022].

4. Análise Comparativa

A Tabela 2 apresenta uma comparação dos estudos selecionados segundo 5 questões de pesquisa centrais: *estratégias de melhoria de qualidade, dimensões e métricas de qualidade, tecnologias, principais desafios e avaliação de resultados*. Elas são detalhadas a seguir.

4.1. Estratégias de melhoria da qualidade dos dados

Os estudos convergem ao reconhecer que a qualidade dos dados deve ser tratada como um processo dinâmico e contínuo [Ehsani-Moghaddam et al. 2021]. Também há consenso quanto à importância do envolvimento de especialistas de domínio na definição de critérios e regras, assegurando relevância clínica e aplicabilidade prática [Diaz-Garelli et al. 2019].

As divergências aparecem principalmente na forma de operacionalização dessas estratégias. Parte da literatura defende maior automação e escalabilidade por meio de técnicas avançadas, enquanto outros trabalhos enfatizam abordagens contextualizadas, com uso de anotações especializadas e curadoria direcionada [Griffiths et al. 2024, Berman et al. 2024, Xie et al. 2025]. Observa-se ainda distinção entre iniciativas voltadas à preparação de bases retrospectivas para pesquisa [Thuraisingam et al. 2021] e propostas de monitoramento contínuo em tempo real [Joshi et al. 2023, Kamal et al. 2025].

Table 2. Análise comparativa dos estudos selecionados

Estudo	Estratégia	Dimensões	Tecnologias	Desafios	Resultados
Benkerourou & Bourouis (2025)	MDM Integrado à governança	Completeness, Unicidade, Consistência	Talend MDM, BPMN	Sistemas fragmentados	-70% duplicidade; retestes $\leq 2\%$
Berman et al. (2024)	Aplicação de DQD <i>framework</i> e checklist para avaliar fenótipos de câncer (DCIS).	Completeness, Conformance, Plausibilidade	OMOP CDM, Python, Jupyter Notebook, OpenEpi.	Divergência EHR	2.209 pacientes validados
Diaz-Garelli et al. (2019)	Padrões especializados	Correção, Completude	MySQL, R, i2b2	Falta de padrões	61 requisitos; 705 problemas detectados
Ehsani-Moghaddam et al. (2021)	<i>Dataset</i> para pesquisa	Acurácia, Completude	SQL, SAS	Dados dispersos	<i>Dataset</i> com 24 variáveis
Fernandes et al. (2022)	Integração via mineração de texto	Padronização diagnóstica	Regex, NLP, SQL	Notas não estruturadas	1.806 pacientes integrados automaticamente
Griffiths et al. (2024)	Etiquetas ontológicas	Cobertura, Contaminação	GenEpiO, Kraken2	Falta de padronização	2.255 sequências avaliadas
Hall et al. (2021)	UDC para limpeza automática	Acurácia	Deep Learning	Dados prejudicando a confiabilidade da IA	Acurácia elevada
Jarmakovic (2025)	ML (detecção de anomalias)	Completeness, Acurácia	scikit-learn, MLflow, Python	Valores ausentes	75,3% de acurácia
Jeon et al. (2023)	Aprendizado Federado	Relevância, Completude	FL, LSTM	Alto custo computacional	Acurácia 87–88%
Joshi et al. (2023)	Detectar outliers	Acurácia, F1	TODS, FlaSH	Alto volume	Superou <i>Deep Learning</i>
Kamal et al. (2025)	Monitoramento <i>ETL</i>	Completeness, taxa de erro	ELK, FHIR	Análise manual	Milhões de arquivos processados
Kim et al. (2021)	5 dimensões e 415 regras	Completeness, validade, acurácia	OMOP CDM	Heterogeneidade EHR	NPR 90–96%
Maity et al. (2024)	Plataforma para fornecer dados	Gradação qualidade	Stack moderna	Dados anotados erroneamente	<i>Dataset</i> padrão-ouro
Marteau et al. (2025)	DQD para atualizar infra de pesquisa	Conformidade OMOP	R Shiny, Power BI	Silos de dados	83,2% aprovação
Monteiro et al. (2025)	Regras ponderadas	Acurácia, Consistência	Great Expectations	Fontes complexas	Escore 75%
Olamiyan & Owoseni (2022)	<i>PyCaret</i> vs OMS	Consistência	<i>PyCaret</i>	Baixo investimento	Melhor detecção
Romero et al. (2022)	Registro ágil	Padronização fluxo	Java, FHIR	Falta padronização	Sistema multiestadual
Sendak et al. (2022)	ML-DQA	Completeness, Plausibilidade	RxNorm API	Dados pouco auditados	2.999 checagens
Shobha Savarimuthu (2021)	CLUSTIMP	RMSE, Acurácia	ART2, EM	Perda de dados	Erro $\downarrow$ 2–11%
Thuraisingam et al. (2021)	Avaliação EHR	Completeness	Stata	IMC ausente (68%)	Preditores inaptos
Wang et al. (2021)	Escore estatístico	Plausibilidade	Código customizado	Limitares rígidos	89% detecção
Xie et al. (2025)	<i>LLM</i> para geração de regras <i>SQL</i>	Completeness, Consistência	<i>GPT</i> , <i>Qwen</i> , <i>SQL</i>	Custo manual alto	97% cobertura; 10x mais rápido

No campo das estratégias emergentes, destaca-se o avanço de soluções automatizadas para geração e validação de regras de qualidade, como modelos generativos capazes de escalar verificações complexas [Xie et al. 2025], *frameworks* estruturados de garantia de qualidade integrados ao ciclo de desenvolvimento de modelos [Sendak et al. 2022] e técnicas de limpeza automática orientadas ao desempenho de IA [Hall et al. 2021]. Também se observam métodos algorítmicos para detecção de anomalias e imputação de dados [Joshi et al. 2023, Shobha and Savarimuthu 2021, Wang et al. 2021], bem como abordagens distribuídas para contextos sensíveis, como o AF [Jeon et al. 2023].

De modo geral, as abordagens variam de intervenções estruturais em governança e reorganização de fluxos até a incorporação de tecnologias de ponta, refletindo uma evolução da área rumo à maior escalabilidade e alinhamento entre qualidade de dados e uso analítico.

4.2. Dimensões e métricas de qualidade

As dimensões de qualidade mais recorrentes incluem *completude* e *acurácia*, que avaliam, respectivamente, a presença de valores ausentes e a correspondência dos dados com a realidade clínica [Ehsani-Moghaddam et al. 2021, Kim et al. 2021, Wang et al. 2021, Benkherourou and Bourouis 2025, Monteiro et al. 2025]. Também são frequentes *consistência* e *conformidade*, voltadas à integridade lógica entre registros e à aderência a padrões estruturais, como o OMOP CDM [Olaniyan and Owoseni 2022, Marteau et al. 2025, Xie et al. 2025]. Já a *plausibilidade* e a *temporalidade* examinam a credibilidade dos valores e a coerência longitudinal dos eventos clínicos [Wang et al. 2021, Sendak et al. 2022, Berman et al. 2024].

Outras dimensões relevantes são *atualidade* e *unicidade*, relacionadas à adequação temporal e à representação única das entidades [Ehsani-Moghaddam et al. 2021, Kim et al. 2021]. Em contextos de IA, destacam-se *reusabilidade*, *relevância* e *valor agregado* dos dados [Jeon et al. 2023, Jarmakovica 2025].

Para mensuração, os estudos empregam indicadores agregados, como *taxas de aprovação* e *escores ponderados por relevância clínica* [Kim et al. 2021] [Monteiro et al. 2025], além de *etiquetas ontológicas* para marcação de problemas específicos [Griffiths et al. 2024] e *métricas estatísticas*, como *RMSE* e *F-score*, para avaliar técnicas de imputação [Shobha and Savarimuthu 2021, Fernandes et al. 2022].

4.3. Tecnologias

As tecnologias identificadas concentram-se principalmente em *soluções de IA*, *arquiteturas de dados* e *ferramentas de validação*. *Modelos de LLM* são empregados para geração automática de regras de qualidade [Xie et al. 2025], enquanto *bibliotecas de AM* em *Python*, como *scikit-learn* e *LightGBM*, apoiam imputação e detecção de anomalias [Jarmakovica 2025] [Olaniyan and Owoseni 2022].

No nível de infraestrutura, destacam-se *arquiteturas baseadas em ELK Stack* e padrão *FHIR* para monitoramento de ingestão de dados em tempo real [Kamal et al. 2025], além de *painéis* como o *DQD* da *OHDSI* para validação de bases no *OMOP CDM* [Kim et al. 2021, Marteau et al. 2025]. *Frameworks* como o *FlaSH* integram modelos estatísticos escaláveis para análise contínua de fluxos [Joshi et al. 2023].

Também são recorrentes *plataformas de MDM* e *ferramentas de modelagem BPMN* [Benkherourou and Bourouis 2025], *ambientes analíticos* como *Databricks* e *APIs REST/FHIR* [Romero et al. 2022], *frameworks de validação* como *Great Expectations* [Monteiro et al. 2025], *APIs clínicas* como *RxNorm* [Sendak et al. 2022] e *ontologias semânticas* como *GenEpiO* em contextos distribuídos [Jeon et al. 2023, Griffiths et al. 2024]. *Softwares estatísticos* como *R*, *SAS* e *Stata* permanecem relevantes para análises retrospectivas e avaliação de plausibilidade [Diaz-Garelli et al. 2019, Ehsani-Moghaddam et al. 2021, Wang et al. 2021].

4.4. Desafios

A implementação de estratégias de qualidade de dados em saúde enfrenta desafios estruturais, técnicos e organizacionais. Destacam-se a *fragmentação e heterogeneidade* dos sistemas de EHR, com dados distribuídos em múltiplas tabelas e frequentemente armazenados como texto livre, exigindo mineração complexa para uso em pesquisa [Ehsani-Moghaddam et al. 2021, Fernandes et al. 2022, Monteiro et al. 2025]. A *falta de padronização terminológica e interoperabilidade*, especialmente no mapeamento para padrões como *LOINC*, *RxNorm* e *OMOP CDM*, gera inconsistências e dificuldades de integração [Kim et al. 2021, Romero et al. 2022, Marteau et al. 2025].

Há ainda um descompasso entre coleta clínica e uso analítico, resultando em *alta incompletude* de variáveis críticas para predição, como *IMC*, comprometendo aplicações de IA [Thuraisingam et al. 2021, Jarmakovica 2025]. Tecnologias emergentes, como *LLMs* e *AF*, ampliam a escala, mas introduzem riscos de *interpretações incorretas*, *alto custo computacional* e desafios de *colaboração institucional* [Xie et al. 2025] [Jeon et al. 2023]. Soma-se a isso a *escassez de governança e competências digitais*, tornando a criação e manutenção de regras de qualidade um processo oneroso e difícil de sustentar em ambientes de saúde pública [Olaniyan and Owoseni 2022, Joshi et al. 2023].

4.5. Avaliação dos resultados

Os estudos demonstram ganhos expressivos em eficiência e precisão na gestão da qualidade de dados. O *LLM-DQR* alcançou até 99,6% de cobertura e reduziu em mais de 10 vezes o tempo de criação de regras [Xie et al. 2025], enquanto técnicas baseadas em *regex* permitiram extração quase perfeita de dados clínicos [Fernandes et al. 2022]. *Frameworks* de MDM reduziram duplicidades em 70% [Benkherourou and Bourouis 2025] e soluções de monitoramento em tempo real garantiram integridade técnica em larga escala [Kamal et al. 2025].

Indicadores padronizados evidenciaram variações relevantes entre instituições. O *DQ4HEALTH* apresentou *NPR* entre 90% e 96% [Kim et al. 2021], o *DQD* atingiu 83,2% de aprovação com falhas em sistemas legados [Marteau et al. 2025], e escores ponderados em COVID-19 variaram de 75% a 98% conforme o grau de padronização [Monteiro et al. 2025]. Métodos de detecção de anomalias, como o *FlaSH*, superaram modelos tradicionais [Joshi et al. 2023] e escores longitudinais apresentaram alto poder discriminatório [Wang et al. 2021]. Outras ferramentas também identificaram inconsistências relevantes em vigilância e bases clínicas [Diaz-Garelli et al. 2019, Olaniyan and Owoseni 2022].

O uso de AM para melhoria de dados mostrou resultados consistentes: técnicas de imputação elevaram a completude e permitiram modelos preditivos robustos

[Shobha and Savarimuthu 2021, Jarmakovica 2025], enquanto *frameworks* de garantia de qualidade evidenciaram a necessidade de limpeza prévia sistemática [Sendak et al. 2022]. Abordagens como UDC aumentaram a acurácia de modelos ao remover dados ruidosos [Hall et al. 2021] e estratégias distribuídas e ágeis apoiaram padronização em contextos complexos [Romero et al. 2022, Jeon et al. 2023, Griffiths et al. 2024]. Ainda assim, a alta proporção de campos críticos ausentes, como IMC, reforça a importância da validação antes da modelagem [Thuraisingam et al. 2021].

5. Conclusão

Este trabalho analisou 22 estudos, entre 2019 e 2025, sobre avaliação e melhoria da qualidade de dados no domínio de Saúde, evidenciando que qualidade deve ser tratada como um processo contínuo e contextual, que combina automação, governança e participação de especialistas. Observou-se convergência nas dimensões clássicas e no uso de indicadores e escores padronizados (*NPR*, *WPR* e métricas ponderadas) para mensuração comparável [Kim et al. 2021, Wang et al. 2021, Berman et al. 2024, Monteiro et al. 2025]. Contudo, mesmo com modelos comuns como o OMOP CDM, persistem variações institucionais, indicando que padronização estrutural não garante consistência semântica [Kim et al. 2021, Marteau et al. 2025].

No eixo tecnológico, há evolução de *dashboards* e *frameworks* de validação para soluções baseadas em AM, monitoramento em tempo real e uso de *LLMs* e AF, que ampliam escala, mas introduzem novos riscos e custos [Jeon et al. 2023, Joshi et al. 2023, Jarmakovica 2025, Kamal et al. 2025, Xie et al. 2025].

Os resultados mostram ganhos relevantes em eficiência, cobertura e redução de retrabalho [Fernandes et al. 2022, Benkherourou and Bourouis 2025, Xie et al. 2025], porém persistem desafios como fragmentação, lacunas terminológicas e alta ausência em variáveis críticas, reforçando a necessidade de validação de adequação antes da modelagem [Thuraisingam et al. 2021, Romero et al. 2022].

Após análise dos estudos selecionados, direções futuras relevantes incluem a construção de arquiteturas que combinem geração assistida e regras criadas por *LLMs*, validação humana estruturada e instrumentos de apoio à decisão, como *dashboards* e logs explicáveis, capazes de guiar especialistas na interpretação de indicadores, no refinamento contínuo de regras e na consolidação de uma governança de qualidade orientada ao uso e ao valor do dado em saúde.

References

- Benkherourou, C. and Bourouis, A. (2025). A framework to enhance data quality in master data management process: a healthcare study. *ITEGAM-JETIA*, 11(55):204–211.
- Berman, L., Ostchega, Y., Giannini, J., Anandan, L. P., Clark, E., Spotnitz, M., Sulieman, L., Volynski, M., and Ramirez, A. (2024). Application of a data quality framework to ductal carcinoma in situ using electronic health record data from the all of us research program. *JCO Clinical Cancer Informatics*, 8:e2400052.
- Diaz-Garelli, J.-F., Bernstam, E. V., Lee, M., Hwang, K. O., Rahbar, M. H., and Johnson, T. R. (2019). Datagauge: A practical process for systematically designing and

- implementing quality assessments of repurposed clinical data. *eGEMs (Generating Evidence & Methods to improve patient outcomes)*, 7(1):32.
- Ehsani-Moghaddam, B. et al. (2021). Data quality in healthcare: a report of practical experience with the canadian primary care sentinel surveillance network. *Health Information Management Journal*, 50(1-2):88–92.
- Fernandes, M., Donahue, M. A., Hoch, D., Cash, S., Zafar, S., Jacobs, C., Hosford, M., Voinescu, P. E., Fureman, B., Buchhalter, J., McGraw, C. M., Westover, M. B., and Moura, L. M. V. R. (2022). A replicable, open-source, data integration method to support national practice-based research & quality improvement systems. *Epilepsy Research*, 186:107013.
- Goldstein, N. D., Kahal, D., Testa, K., Gracely, E. J., and Burstyn, I. (2022). Data quality in electronic health record research: An approach for validation and quantitative bias analysis for imperfectly ascertained health outcomes via diagnostic codes. *Harvard Data Science Review*, (4.2).
- Griffiths, E. J. et al. (2024). Pha4ge quality control contextual data tags: standardized annotations for sharing public health sequence datasets with known quality issues to facilitate testing and training. *Microbial Genomics*, 10:001260.
- Hall, J. M. M., Pham, D., Nguyen, T. V., VerMilyea, M., Lynn, R., Diakiw, S. M., and Perugini, M. (2021). Untrainable data cleansing for ai in healthcare. *Scientific Reports*, 11.
- Jarmakovica, A. (2025). Machine learning-based strategies for improving healthcare data quality: an evaluation of accuracy, completeness, and reusability. *Frontiers in Artificial Intelligence*, 8:1621514.
- Jeon, K.-C., Han, G.-S., Han, C.-Y., and Chong, I. (2023). Federated learning model for contextual sensitive data quality applications: Healthcare use case. In *SmartCore; Hankuk University of Foreign Studies*. IEEE.
- Joshi, A., Mazaitis, K., Rosenfeld, R., and Wilder, B. (2023). Computationally assisted quality control for public health data streams. *Carnegie Mellon University*.
- Kaloyanova, K., Naydenova, I., and Kovacheva, Z. (2021). Addressing data quality in healthcare. In *CEUR Workshop Proceedings*, volume 2933, pages 156–164.
- Kamal, M. M., Kutafina, E., and Beyan, O. (2025). Real-time automated monitoring architecture for ensuring quality etl processes in healthcare data: A case study from cologne university hospital’s medical data integration center. In *Intelligent Health Systems*, pages 192–197. IOS Press.
- Kim, K.-H., Choi, W., Ko, S.-J., Chang, D.-J., Chung, Y.-W., Chang, S.-H., Kim, J.-K., Kim, D.-J., and Choi, I.-Y. (2021). Multi-center healthcare data quality measurement model and assessment using omop cdm. *Applied Sciences*, 11(19):9188.
- Kitchenham, B. (2004). *Procedures for Performing Systematic Reviews*. Keele University.
- Maity, D., Satish, R., Jadeja, D. A., Dharmaraju, R., Chandru, V., Sundaresan, R., Singh, H., and Pal, D. (2024). Midas: a new platform for quality-graded health data for ai-enabled healthcare in india. *Nature Medicine*.

- Marteau, B. L., Hornback, A., Zhong, Y., Lowson, C., Woloff, J., Smith, B. M., Hilton, C., and Wang, M. D. (2025). Improving a large healthcare system research data warehouse using ohdsi's data quality dashboard. In *Intelligent Health Systems*. Georgia Institute of Technology; Shriners Hospitals for Children.
- Monteiro, S., Oliveira, B., Elvas, L., and Ferreira, J. (2025). Improving healthcare data quality: A case study in a portuguese hospital. *Procedia Computer Science*, 256:1030–1038.
- Nasir, W. M. H. M., Abdullah, R., Jusoh, Y. Y., and Abdullah, S. (2022). Big data analytics quality in enhancing healthcare organizational performance: A conceptual model development. *International Journal of Advanced Computer Science and Applications*, 13(11).
- Olaniyan, F. M. A. and Owoseni, A. (2022). Toward improved data quality in public health: Analysis of anomaly detection tools applied to hiv/aids data in africa. In *IST-Africa 2022 Conference Proceedings*.
- Romero, L., Carneiro, P. B., Riley, C., Clark, H., Uy, R., Park, M., Mawokomatanda, T., Bombard, J. M., Hinckley, A., and Skapik, J. (2022). Building capacity of community health centers to overcome data challenges with the development of an agile covid-19 public health registry: a multistate quality improvement effort. *Journal of the American Medical Informatics Association*, 29(1):80–88.
- Sendak, M. et al. (2022). Development and validation of ml-dqa – a machine learning data quality assurance framework for healthcare. *Proceedings of Machine Learning Research*, 182:1–19.
- Shahnawaz, M. and Kumar, M. (2025). A comprehensive survey on big data analytics: characteristics, tools and techniques. *ACM Computing Surveys*, 57(8):1–33.
- Shobha, K. and Savarimuthu, N. (2021). Clustering based imputation algorithm using unsupervised neural network for enhancing the quality of healthcare data. *Journal of Ambient Intelligence and Humanized Computing*, 12:1771–1781.
- Thuraisingam, S., Manski-Nankervis, J.-A., Choong, P., and Dowsey, M. (2021). Suitability of general practice electronic health records for clinical prediction model development. *BMC Medical Informatics and Decision Making*, 21:297.
- Wang, H., Belitskaya-Levy, I., Wu, F., Lee, J. S., Shih, M.-C., Tsao, P. S., and Lu, Y. (2021). A statistical quality assessment method for longitudinal observations in electronic health record data with an application to the va million veteran program. *BMC Medical Informatics and Decision Making*, 21:289.
- Xie, S., Cai, H., Sun, Y., and Lv, X. (2025). Llm-dqr: Large language model-based automated generation of data quality rules for electronic health records. *Journal of Biomedical Informatics*, 172:104951.